UNIVERSIDAD CARLOS III DE MADRID

DEPARTAMENTO DE INGENIERÍA TELEMÁTICA

TESIS DOCTORAL

Convergencia de Tecnologías Ópticas y Ethernet en LAN, MAN y SAN: Nuevas Arquitecturas, Análisis de Prestaciones y Eficiencia Energética

Autor: **Gerson Rodríguez de los Santos López**
Ingeniero de Telecomunicación, Máster en Ingeniería Telemática

Director: **Prof. Dr. José Alberto Hernández Gutiérrez**
Doctor en Ciencias de la Computación

Leganés, Junio de 2014

UNIVERSIDAD CARLOS III DE MADRID

DEPARTMENT OF TELEMATIC ENGINEERING

PhD THESIS

Converged Optical Ethernet Technologies in the LAN, MAN and SAN: Novel Architectures, Performance Analysis and Energy Efficiency

Author: **Gerson Rodríguez de los Santos López, MsC**

Supervisor: **Prof. Dr. José Alberto Hernández Gutiérrez**
Doctor in Computer Science

Leganés, June, 2014

CONVERGENCIA DE TECNOLOGÍAS ÓPTICAS Y ETHERNET EN LAN, MAN y SAN: NUEVAS ARQUITECTURAS, ANÁLISIS DE PRESTACIONES Y EFICIENCIA ENERGÉTICA

CONVERGED OPTICAL ETHERNET TECHNOLOGIES IN THE LAN, MAN AND SAN: NOVEL ARCHITECTURES, PERFORMANCE ANALYSIS AND ENERGY EFFICIENCY

Autor: Gerson Rodríguez de los Santos López
Director: Prof. Dr. José Alberto Hernández Gutiérrez

Tribunal nombrado por el Mgfco. y Excmo. Sr. Rector de la Universidad Carlos III de Madrid, el día ___ de _____ de _____.

Firma del tribunal calificador:

Firma:

Presidente:

Vocal:

Secretario:

Calificación:

Leganés, ___ de _____ de _____.

A mi abuelo Artemio, que siempre estuvo orgulloso de mí y por pocos meses no pudo estar aquí para ver esto.
A mi tía Constanza, que tampoco está hoy con nosotros.
A mi madre, Noemi, por quererme incodicionalmente.
A mi padre, Ángel, por entenderme siempre.
A mi abuela, Angelita, por enseñarme la importancia del cordero en el cocido.
A mi tío Rafael, por despertar la inquietud científica en mí cuando era pequeño.
A mi tío Luis, por enseñarme la importancia de la política.
A la mejor de mis tías, Isabelita, por ser la más mejor de las mejores.
A mi tío Josué, por ser un apoyo en mis creencias.
Y a mis primos en su conjunto, que son muchos para mencionar y no me cabrían en una página, pero que los quiero mucho igualmente.

Many that live deserve death. And some that die deserve life. Can you give it to them? Then do not be too eager to deal out death in judgement. For even the very wise cannot see all ends.

John Ronald Reuel Tolkien. *The Lord of The Rings.*

Agradecimientos

Quisiera agradecer en estas líneas a todas aquellas personas que han estado junto a mí y sin las cuales no hubiera sido posible terminar esta tesis. Por un lado, a mi familia, sin cuyo apoyo no hubiera podido ya no sólo terminar esta tesis, sino ser quien soy hoy.

A mis amigos, especialmente Alejandro Jiménez, Maximino Arévalo y Roberto Marines, que no solamente me apoyaron moralmente sino que también me dieron su punto de vista en temas técnicos de la tesis en los campos en que ellos son expertos.

Y, finalmente, a mi tutor, José Alberto Hernández y al que me introdujo en el mundo de la investigación, David Larrabeiti, por abrirme ambos un mundo nuevo.

A todos ellos, gracias de corazón.

Abstract

The development of Information Technologies in the last decades, especially the last two, together with the introduction of computing devices to the mainstream consumer market, has had the logical consequence of the generalisation of the Internet access. The explosive development of the smartphone market has brought ubiquity to that generalisation, to the point that social interaction, content sharing and content production happens all the time. Social networks have all but increased that trend, maximising the diffusion of multimedia content: images, audio and video, which require high network capacities to be enjoyed quickly.

This need for endless bandwidth and speed in information sharing brings challenges that affect mainly optical Metropolitan Area Networks (MANs) and Wide Area Networks (WANs). Furthermore, the wide spreading of Ethernet technologies has also brought the possibility to achieve economies of scale by either extending the reach of Ethernet Local Area Networks (LANs) to the MAN and WAN environment or even integrating them with Storage Area Networks (SANs). Finally, this generalisation of telecommunication technologies in every day life has as a consequence an important rise in energy consumption as well. Because of this, providing energy efficient strategies in networking is key to ensure the scalability of the whole Internet.

In this thesis, the main technologies in all the fields mentioned above are reviewed, its core challenges identified and several contributions beyond the state of the art are suggested to improve today's MANs and WANs. In the first contribution of this thesism, the integration between Metro Ethernet and Wavelength Division Multiplexion (WDM) optical transparent rings is explored by proposing an adaptation architecture to provide efficient broadcast and multicast. The second contribution explores the fusion between transparent WDM and OCDMA architectures to simplify medium access in a ring.

Regarding SANs, the third contribution explores the challenges in SANs through the problems of Fibre Channel over Ethernet due to buffer design issues. In this contribution, analysis, design and validation with FCoE traces and simulation is provided to calculate buffer overflow probabilities in the absence of flow control mechanisms taking into account the bursty nature of SAN traffic.

Finally, the fourth and last contribution addresses the problems of energy efficiency in Plastic Optical Fibres (POF), a new kind of optical fibre more suitable for transmission in vehicles and for home networking. This contribution suggests two packet coalescing strategies to further improve the energy effiency mechanisms in POFs.

Resumen

El desarrollo de las Tecnologías de la Información en las últimas décadas, especialmente las últimas dos, junto con la introducción de dispositivos informáticos al mercado de masas, ha tenido como consecuencia lógica la generalización del acceso a Internet. El explosivo desarrollo del mercado de teléfonos inteligentes ha añadido un factor de ubicuidad a tal generalización, al extremo de que la interacción social, la compartición y produción de contenidos sucede a cada instante. Las redes sociales no han hecho sino incrementar tal tendencia, maximizando la difusión de contenido multimedia: imágenes, audio y vídeo, los cuales requieren gran capacidad en las redes para poder obtenerse con rapidez.

Esta necesidad de ancho de banda ilimitado y velocidad en la compartición de información trae consigo retos que afectan principalmente a las Redes de Área Metropolitana (Metropolitan Area Networks, MANs) y Redes de Área Extensa (Wide Area Networks, WANs). Además, la gran difusión de las tecnologías Ethernet ha traído la posibilidad de alcanzar economías de escala bien extendiendo el alcance de Ethernet más allá de las Redes de Área Local (Local Area Networks, LANs) al entorno de las MAN y las WAN o incluso integrándolas con Redes de Almacenamiento (Storage Area Networks, SANs). Finalmente, esta generalización de las tecnologías de la comunicación en la vida cotidiana tiene también como consecuencia un importante aumento en el consumo de energía. Por tanto, desarrollar estrategias de transmisión en red eficientes energéticamente es clave para asegurar la escalabilidad de Internet.

En esta tesis, las principales tecnologías de todos los campos mencionados arriba serán estudiadas, sus más importantes retos identificados y se sugieren varias contribuciones más allá del actual estado del arte para mejorar las actuales MANs y WANs. En la primera contribución de esta tesis, se explora la integración entre Metro Ethernet y anillos ópticos transparentes por Multiplexión en Longitud de Onda (Wavelength Division Multiplex, WDM) mediante la proposición de una arquitectura de adaptación para permitir la difusión y multidifusión eficiente. La segunda contribución explora la fusión entre las arquitecturas transparentes WDM y arquitecturas por Accesso Dividido Múltiple por Códigos Ópticos (OCDMA) para simplificar el acceso en una red en anillo.

En lo referente a las SANs, la tercera contribución explora los retos en SANs a través de los problemas de Fibre Channel sobre Ethernet debido a los problemas en el diseño de búferes. En esta contribución, se provee un análisis, diseño y validación con trazas FCoE para calcular las probabilidades de desbordamiento de buffer en ausencia de mecanismos de control de flujo teniendo en cuenta la naturaleza rafagosa del tráfico de SAN.

Finalmente, la cuarta y última contribución aborda los problemas de eficiencia energética en Fibras Ópticas Plásticas (POF), una nueva variedad de fibra óptica más adecuada para la

transmisión en vehículos y para entornos de red caseros. Esta contribución sugiere dos estrategias de agrupamiento de paquetes para mejorar los mecanismos de eficiencia energética en POFs.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Motivation of this thesis

The development of Information Technologies in the last decades, especially the last two, together with the introduction of computing devices to the mainstream consumer market, has had the consequence of the generalisation of the Internet access as well. The explosive development of the smartphone market has brought ubiquity to that generalisation, to the point that social interaction, content sharing and content production happens all the time. Social networks have all but increased that trend, maximising the diffusion of multimedia content: images, audio and video, which require high network capacities to be enjoyed quickly. A side effect of this capacity increase is that there is less importance today in storing that information since it can be recoverable in the collective memory of the Internet. That is, there is usually the possibility of recovering a copy of a piece of information from the same place from where it was initially obtained. It is way easier to provide a link to a content provider than transmitting individual copies to each person connected to a network, more so given the growth of content sharing providers (Youtube) and the Content Distribution Networks (Eg: Akamai). Another side effect of this model of distribution is that content providers require massive datacentres to give the intended service.

But all the above mentioned is a product, rather than a cause, of the development of computing and telecommunication standards. Contrary to popular perception, it has been pointed out that it is not the need for services that brings the increase in bandwidth but it is rather the increase in bandwidth that fosters the creation of new services that use that bandwidth. With such an increase comes the need for new architectures in Metropolitan Area Networks (MANs) and Wide Area Networks (WANs), since that growth brings new problems to the table.

One of these problems is the electronic bottleneck, that is, the inability of electronic devices to switch back all the optical traffic to the electric domain in every point of the network. Link capacity is only part of the transmission problem. All the traffic transmitted through a set of links has to be switched at some point in the network. There is no use in having high link capacities without switching devices capable of absorbing all (or most of) the traffic in those links. In this sense, transparent optical networks are a way of avoiding this electronic bottleneck. This issue has become pretty evident with the introduction of optical networks populated by a myriad of parallel channels. There are two technologies

aimed at providing several parallel transmission channels in a network, namely Wavelength Division Multiplex (WDM) systems and Optical Code Division Multiple Access (OCDMA) networks.

Another problem posed by the increase in bandwidth is the need for integration among different technologies. Since its standardisation in the 1980s, Ethernet has become the reference standard for Local Area Networks (LAN) due to its simplicity, plug-and-play nature and cost-effectiveness. This has led the IEEE to broaden its scope to the Metropolitan Area Networks (MAN). In the course of its development, the Ethernet is bound to meet with optical transparent technologies, which leads to the challenge of integration of Ethernet with these optical network architectures. There is also a challenge in integrating several optical network architectures to provide different ways to access the medium.

However, the development of Ethernet does not only meet with optical technologies. For quite some time, storage technologies have run parallel to transmission technologies, as two separate worlds. In the last two decades, though, the need for concurrent access to information has lead to the development of technologies to access storage media through a data switched network. There are two main paradigms to provide access to storage through an Ethernet network. One is Network Attached Storage (NAS), which uses a common packet network as a means of accessing data, and the other one is Storage Area Networks (SANs), which by contrast are networks specifically engineered to provide access to stored data. One of the first technologies proposed for SANs was Fibre Channel (FC). Fibre Channel, though, needs its own network, which does not suit for economy of scale purposes. Given the omnipresence of Ethernet in LANs and consequently in datacentres, Fibre Channel over Ethernet (FCoE) is the natural evolution to provide storage access through a network. Unfortunately, the nature of both technologies, Fibre Channel on one side and Ethernet on the other, could not be more different. While FC is designed to provide reliable transmission on a dedicated network in short distances, Ethernet networks are usually suited to provide non-reliable access for longer distances. While the reliability of Ethernet has been recently improved by the Data Centre Bridging (DCB) group, there is still a challenge derived from the longer distances that can be found in Ethernet networks as of today. This problem is accentuated by the growing size of data centres and the need for producing instant or almost instant backups in a redundant datacentre away from the primary one to protect corporate information from disasters. This last point implies the jump of Storage Area Networks to the MAN.

Finally, an increase of activity brings associated, of course, an increase in energy consumption. While energy has historically not been the main preoccupation of networking research, it has become so in the last decade precisely because of the growing volume of networking activity. One of the many areas in which energy efficiency has been researched is in the field of transmission because of the needlessness of having transceivers constantly at full power even in low load conditions. A clear challenge in making the energy consumption as linear to the link load as possible can be identified here. Although Energy Efficiency has been clearly addressed in Ethernet, there is still work to be done in a part of the optical technologies field that has not been addressed until recently by the VDE-0885-763-1 standard for high-speed transmission over Plastic Optical Fibres (POFs).

All the aforementioned challenges affect mainly optical MANs and WANs, which are the main focus of this thesis. As it can be deduced from the paragraphs above, this implies

the need to study different but related technologies in the networking field in order to actually solve these challenges. The main technologies involved in this thesis can be seen in Fig. 1.1. There are three main areas in which to classify them:

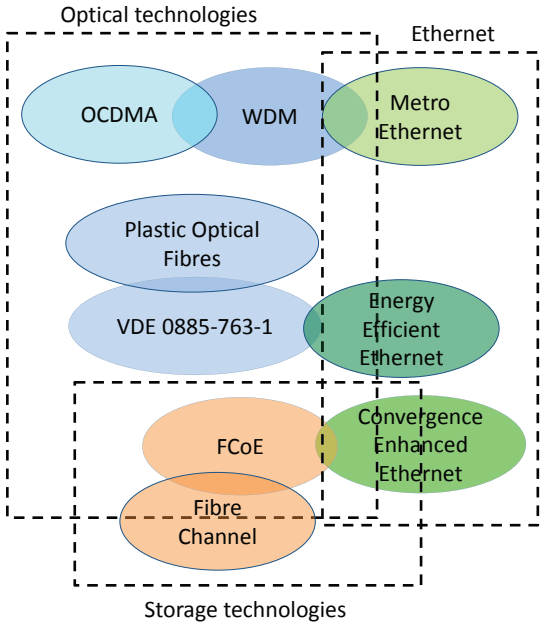


Figure 1.1: Main technological areas covered in this thesis.

- Optical Technologies: This area involves transparent Optical Network architectures, like WDM and OCDMA. Integration betweeen those heterogeneous technologies is also studied. Plastic Optical Fibres are also investigated along with the recent VDE-0885-763-1 standard for high-speed transmission over Plastic Optical Fibres.

- Ethernet: The three main threads of Ethernet extension in this thesis are: MAN and WAN extension, covered through amendments to the 802.1 standards; Energy Efficiency awareness, which is covered by Energy Efficient Ethernet; and transmission reliability, which is covered by Convergence Enhanced Ethernet. Due to the ubiquity of Ethernet in current LANs, integration of Ethernet with optical technologies is also necessary. In this sense, Energy Efficient Ethernet (EEE) has inspired part of the contributions of this thesis.

- Storage Technologies: They have been the focus of recent research in the field of communication networks. The ever increasing need for ubiquitous access to information has led to an important effort of integration between storage and network technologies. The result of this is the creation of Storage Area Networks and the fusion of such networks with Ethernet through the Convergence Enhanced Ethernet framework, whose aim is to reduce economic costs and avoid redundant infrastracture, giving as a result the definition of FCoE. Unfortunately, buffer design seems to pose important if not essential limitations to the reach of these networks.

In the next sections, a brief historical evolution of these three main technological areas is provided. Next, a summary of the current state of the art in these three main areas is provided. Finally, the main contributions and their relationships to these challenges and the technologies involved will be presented at the end of this chapter.

1.2 Historical evolution of the technologies studied in this thesis

A timeline for the main technology lines displayed in Fig. 1.1 can be seen in Fig. 1.2. The main milestones shown in Fig. 1.2 are quickly outlined in the next sections.

1.2.1 Ethernet and Metro Ethernet

From the very beginning of its existence, Ethernet was conceived as a technology with a simple MAC level, something which made it a cost effective, plug-and-play technology. Ethernet in its first definition was designed for a bus medium, which required stations to listen to the medium before trasmitting to avoid collisions (Carrier Sense Multiple Access/Collision Detection (CSMA/CD)). Later evolutions of the Ethernet standard were released to incorporate the possibility of separating collision domains, to the point that Ethernet is today a full duplex, collision-less technology that has nothing to do with its first definitions.

Ethernet was developed in 1973 by a group of Xerox engineers. The first experimental Ethernet ran at a speed of 3 Mbps. During the 1980s, the standardisation process began in the IEEE group 802.3, bearing the first Ethernet standard in 1982. The first version of the standard defined Ethernet for thick coaxial cable at a rate of 10 Mbps. Later versions would define, during the 80s and the first half of the 90s, supposed a significant change in the binary rate of Ethernet, jumping from 10 Mbps to 100 Mbps and receiving the name of ”Fast Ethernet”. From 1995 to 1998, Fast Ethernet was defined for several kinds of twisted pair and optical fibre media. But Fast Ethernet was outdated before long when in 1998 the 802.3z standard presented Ethernet at a 1 Gbps rate (Gigabit Ethernet). In contrast to previous Ethernet releases, the first Gigabit Ethernet standard was defined for a optic fibre medium instead of a copper medium. This would also apply to the 10 Gigabit Ethernet standard, 802.3ae in 2002. Despite this, in both cases there was a definition of the 1 Gbps and 10 Gbps standards for twisted pair in 1999 and 2006 respectively. The first 40 Gbps and 100 Gbps standard would arrive in 2010 with the 802.3ba standard. In 2011, a standard for a single wavelength 40G Ethernet arrived through [5]. In general, most 40G and 100G Ethernet specifications for optical fibre use WDM or several fibres to provide the desired binary rate.

The specifications indicated above are those related to the ethernet physical layer, but Ethernet has also a link layer, which is standardised by the IEEE 802.1 group. The 802.1 group also covers bridging and the different versions of the Spanning Tree Protocol (STP). The first IEEE standard of STP appeared in 1990 with 802.1D. STP had a convergence time between 30 to 50 seconds. The standard for the Rapid Spanning Tree Protocol (RSTP) appeared in 2001 with the 802.1w standard and improved this convergence speed to around 6 seconds. RSTP was closely followed by the Multiple Spanning Tree Protocol (MSTP) in 2002, defined in the 802.1s specification, which enabled Spanning Tree in a context with several Virtual Local Area Networks (VLANs). The latest advancement in switching con-

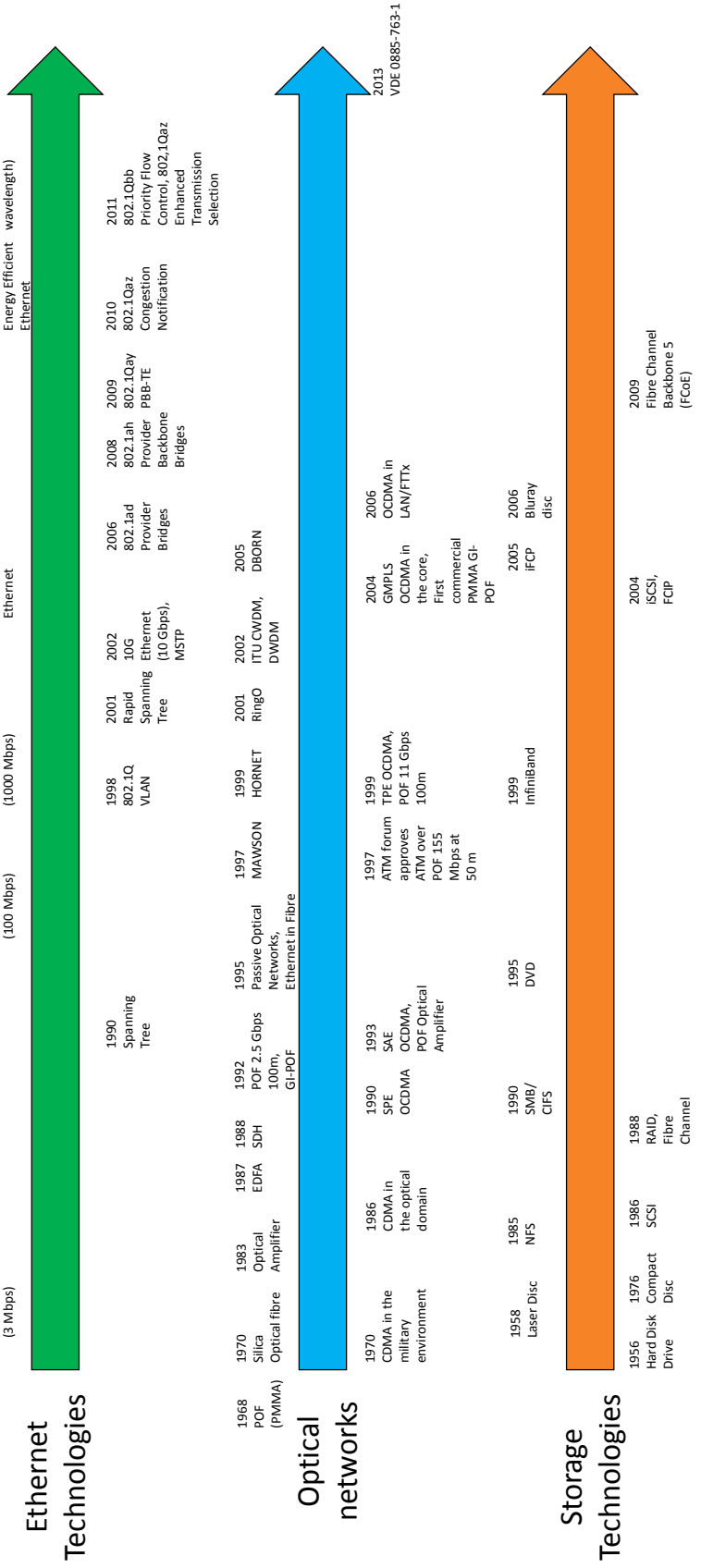


Figure 1.2: Historical evolution of the technologies in this thesis

figuration in Ethernet networks comes from the 802.1aq standard, which defines Shortest Path Bridging (SPB) in Ethernet networks. SPB replaces the old spanning tree protocols by changing the bridging paradigm in Ethernet from a root-centered tree to a shortest path path bridge from each node. It also provides faster convergence and multipath briding. The SPB standard saw the light of day in 2012.

The 802.1Q standard defined the concept of Virtual Local Area Network (VLAN). 802.1Q allows several virtual LANs to share the same physical medium transparently. It appeared in 1998. Further modifications to the switching layer of Ethernet were introduced to make Ethernet scale to large scenarios. These modifications to the 802.1 standard are the Provider Bridges (PB, 802.1ad, 2006), the Provider Backbone Bridges (PBB, 802.1ah, 2008) and the Provider Backbone Bridges Traffic Engineering (PBB-TE, 802.1Qay, 2009) amendments.

Several other extensions of Ethernet have been defined. One of the most important ones lately has been the set of standards brought by the Data Centre Bridging Group (DCB), whose main target has been to enable Ethernet for carrying other kinds of frames which require are more reliable network. The work of this task force gave as a result the 802.1Qaz-2010 (Congestion Notification (CN)), 802.1Qbb-2011 (Priority Flow Control (PFC)) and 802.1Qaz-2011 (Enhanced Transmission Selection (ETS)).

Finally, the 802.3af-2003 and 802.3az-2010 standards brought Power over Ethernet (POE) and Energy Efficient Ethernet (EEE) respectively. Power over Ethernet enables bridges to act as power sources, while the Energy Efficient Ethernet makes several changes to the Ethernet transmission level to make it more energy efficient.

Some of the latest standards, especially the MAN/WAN extensions, Energy Efficient Ethernet and Datacentre Bridging Group extensions will be the subject of review in the state of the art.

1.2.2 Optical technologies

The optical technology field has traditionally moved around two main objectives: to achieve longer reach and higher bandwidth.

Optical technologies have been under study for several centuries, optical fibres included, but it was not until the 1970s that this field boomed due to the invention of the silica optical fibre. While it initially had high attenuation, subsequent advancements reduced it significantly, leading to values so low as those of Standard Single Mode Fibre (SSMF), which has 0.2 dB/km. While this might seem pretty natural to us nowadays, this was not the case when optical fibres were first manufactured, having attenuations of hundreds of dB/km. Another decade would have to pass until the crucial invention of Optical Amplifiers in 1983 at Stanford University and one of their most popular kinds, the Erbium Doped Fibre Amplifiers (EDFA) in 1987.

These advancements in the 1980s led to the standardisation of the Synchronous Optical Networking and Synchronous Digital Hierarchy (SONET/SDH), one of the most popular Time Division Multiplex (TDM) technologies for the MAN among operators. Parallel to this, CDMA gave the jump to the optical domain. In the mid 1990s, optical technologies entered the last mile with Passive Optical Networks. Before that, optical technologies had been mainly conceived for the WAN due to their long reach and high capacity, despite their high

cost. WDM technologies appeared at the end of the 1990s and the beginning of the 2000s, parallel to similar advancements in OCDMA. The first Coarse WDM standard (CWDM) appeared in 2002, with the ITU-T G.694.2 specification. Dense WDM systems were also standardised in 2002 through the ITU-T G.694.1 specification. It is also worth mentioning that, while optical fibre was initially a technology for the MAN and WAN, it has recnetly reached the last mile of telecommunication networks.

Although all the previous progress in optical transmission is built on top of the silica core fibre, there are other types of fibre that have been the subject of research. Plastic Optical Fibres (POF) were conceived in 1968, but unfortunately they were shadowed by the appearance of the silica core fibre in the 1970s, despite their lower fragility. Due to the appeal of silica optical fibres for long reach communications, Plastic Optical Fibres were put aside until the 1990s, when significant milestones regarding the reduction of attenuation in POFs were achieved. At the beginning of the 1990s, significant bit rates were reached POFs, as well as the invention of Plastic Optical Fibre amplifiers in 1993. In 1997, the Asynchronous Transfer Mode (ATM) forum standardised ATM at 155 Mbps over Plastic Optical Fibres and only two years later, 11 Gbps with a reach of 100 m was achieved in POF.

In 2004, the first Poly Methyl Methacrylate (PMMA) Graded Index POFs (PMMA-GI POF) were introduced to the market. The most recent advance in POF has come from the VDE 0885-763-1 standard, which specifies the physical media for high-speed transmission over POF. This standard has been supported mainly by car manufacturers, although home networking environments are also possible.

1.2.3 Storage Technologies

Storage technologies experienced an earlier boom due to the development of computer science around the 1940s and 50s. In 1956, the first hard disks were invented. The advancements in this field continued with the invention of optical discs (Laser disc and compact disc) through Sony and Panasonic mainly. Significant advancements in computer storage came with network file systems in the mid 1980s as well as the appearance of the Small Computer System Interface (SCSI) protocol. Two years later, another crucial technology for storage, the Redundant Array of Independent Disks (RAID) technology would appear. That same year, in 1988, the first technology for Storage Area Networks, Fibre Channel, was born.

In the 1990s the optical discs would recover protagonism through the next step in their evolution, the Digital Versatile Disc (DVD) in 1995. Between the end of that decade and the beginning of the next, several other Storage Area Network technologies appeared, among which the most relevant were probably Internet SCSI (iSCSI) and InfiniBand. In the 2000s, the main advances in storage have come from iSCSI, the Bluray disc and the adaptation of Fibre Channel to Ethernet through the INCITS T11's FC-BB-5 specification, leading to what is known as Fibre Channel over Ethernet (FCoE).

In the next sections, a detailed state of the art of all the technologies mentioned in this section is provided.

1.3 State of the art of the main technologies in this thesis

1.3.1 Ethernet technologies and extensions

As previously mentioned, Ethernet became the preferred technology for the LAN environment due to its initial easier MAC layer implementation. This made Ethernet a cost effective, plug-and-play technology, something that helped it become quickly mainstream, although it still had certain shortcomings. Later on, several changes in the standard eliminated many of its shortcomings, like collisions due to the CSMA/CD medium access control method.

The generalisation of Ethernet in the LAN has brought to the table one interesting possibility: because LANs are Ethernet, it is easier to integrate communication among LANs, which are usually at the endpoints of the Internet, if the Metropolitan and Wide Area Networks are also Ethernet capable. As a result, several converging efforts aimed at making Ethernet a technology suitable for Wide Area Networks have taken place at different organisations. These changes are covered in the majority of this section. A brief review of energy efficiency extensions for Ethernet is made at the end of this section, as it will also be relevant to another contribution in this thesis.

Contributions of the Metro Ethernet Forum

One of the main groups that has worked towards the standardisation of Ethernet services in MANs and WANs is the Metro Ethernet Forum (MEF). In a MAN or WAN, regarding connectivity, from the point of view of the customer, there are three kinds of services required to be provisioned according to the definitions made by the Metro Ethernet Forum [6–8]. In their standards, the MEF defines three types of Ethernet connections to be provided through a MAN or WAN:

- Point-to-point: E-Lines.
- Point-to-multipoint: E-Trees.
- Multipoint-to-multipoint, called E-LAN.

The Metro Ethernet Forum has defined the so called Ethernet Virtual Connections (EVCs) [8], which are associations between User Network Interfaces (UNIs). A UNI is the physical point in which the responsibility of the Service Provider and the responsibility of the Subscriber are divided [8].

The E-Line services are based on point-to-point connections of Ethernet islands. An illustration of it can be seen in Fig. 1.3. In its simplest definition, E-line services can provide symmetrical bandwidth with no QoS (best effort service) between two User Network Interfaces (UNIs). More sophisticated definitions might include QoS terms, like CIR (Commited Information Rate), Commited Burst Size (CBS), Excess Information Rate (EIR) and Excess Burst Size (EBS), guaranteed delay, delay jitter, loss of frames, etc. Multiple E-line services can be defined on the same UNI interface, allowing service multiplexion on the same physical port.

The E-Tree services define a point-to-multipoint connection type. An example of such a service can be seen in Fig. 1.4. In its simplest form, there is a single Root UNI which

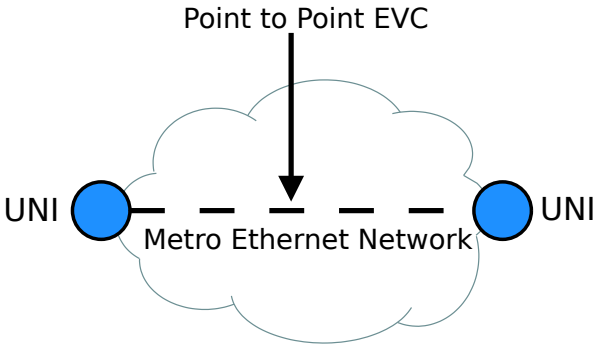


Figure 1.3: E-line service example.

communicates with multiple Leaf UNIs and the Leaf UNIs can only communicate only with the Root UNI. Communication among Leaf UNIs is not allowed. This service is conceived for Internet Access, Voice over IP setups and multicast/broadcast video. In a more complex service scenario, two or more Root UNIs might be supported. As in the previous scenario, each Leaf UNI can only communiate with only the Root UNIs. In this service scenario, a redundant access to the Root UNI is provided.

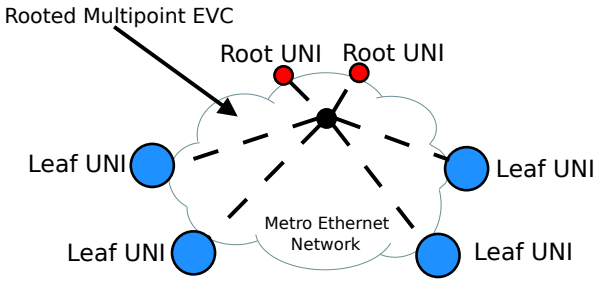


Figure 1.4: E-tree service example.

Finally, E-LAN services define a multipoint-to-multipoint EVC scenario. This services provide a full Ethernet LAN interconnection through a MAN/WAN. An illustration of the concept of E-LAN services can be seen in Fig. 1.5. In its simplest implementation, an E-LAN Service can provide a best effort service among the UNIs. In more complex forms, an E-LAN service may offer certain CIR, CBS, EIR, EBS, delay, jitter and loss for a given Class of Service. Service multiplexion can happen at none, one or more than one UNI. This may lead to the coexistence of E-LAN services with, for instance, E-Line services in the same UNI. The E-LAN service could be used to interconnect a site with other subscriber sites while the E-Line service provided by multiplexion at the same UNI could be used to provide Internet Access.

To provide these services, several modifications to Ethernet have been proposed by the IEEE to improve scalability: Virtual LANs, Provider Bridges, Provider Backbone Bridges and Provider Backbone Bridges - Traffic Engineering. It is worth mentioning that the main three types of Ethernet services mentioned above can also be provided by non-native technologies, like Virtual Private Line Services (VPLS), based on MPLS. This VPLS services provide a layer 2 service across a backbone network [9, 10]. In VPLS, the network emulates

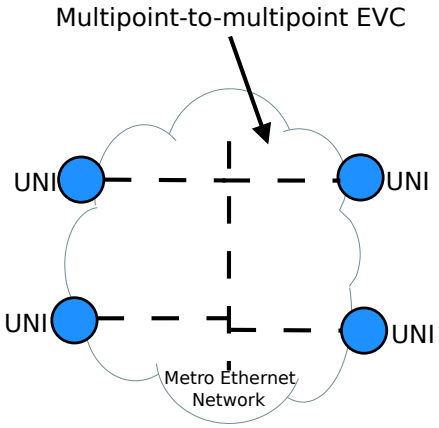


Figure 1.5: E-LAN service example.

the function of an Ethernet bridge. There are also layer 3 based services for private networking [11], but those services are not necessarily Ethernet oriented. In any case, Multi Protocol Label Switching (MPLS) based technologies are out of the scope of this thesis and only the native Ethernet ones will be reviewed.

Introduction to the IEEE standards for Ethernet Metropolitan Area Network extensions

Regarding LAN and MAN architecture, internetworking between LANs and network management, among other issues, there is also a working group in the IEEE, the 802.1. To the scope of this chapter, the first relevant version of the standard is the 802.1Q release, in 1998, which bore the Ethernet tagged VLANs (Virtual LANs). Tagged VLANs allow the multiplexion of several LANs over the same physical medium by simply adding a 12-bit label in the middle of the standard Ethernet header. In this sense, 802.1Q was an important advance for corporate networks, as it provided a way of creating several paralell LANs between departments.

As years have gone by, there has been an increasing movement advocating for the extension of Ethernet to the MAN (Metropolitan Area Networks) and WAN (Wide Area Networks), due to its simplicity and wide adoption in the LAN segments. With the original definition of the 802.1Q in 1998 and the later revision in 2005, it soon became clear that there would be problems with scalability in Ethernet. Its main drawback being the 12 bit VLAN labels, 802.1Q was soon replaced by the 802.1ad (Provider Bridges Standard). Commonly known as Q-in-Q because of the addition of another VLAN-Tag outside of the origigiinal 802.1Q Tag, it provided a greater level of scalability to Ethernet. Nevertheless, this level was not sufficient and, furthermore, not only did it lack scalability for large network scenarios (802.1ad allowed 12 bit tags), but also it lacked service instance differentiation mechanisms as well as other network management features.

The next step would be provided by 802.1ah in 2008. The 802.1ah version defines Provider Backbone Bridges, which highly improves scalability thanks to MAC in MAC forwarding (Ethernet frames inside Ethernet frames), service instance differentiation and client LAN protocols isolation. The 802.1ah is the first Ethernet standard thought for a backbone

network. It allows a high number of service instances (up to 16 million approx.) and forwarding inside the core network keeping forwarding tables at a reasonable size, thanks to the MAC in MAC encapsulation. This encapsulation also prevents backbone bridges from knowing details about client LANs and end stations, improving the scalability of the network.

Even taking all the facts mentioned above into consideration, there were still some features Ethernet needed to become a proper WAN technology. Because Ethernet was thought in its origins to be a simple technology even in terms of forwarding. There are three important features of the Ethernet forwarding mechanism:

- Switching paths are backward-learned.
- There are no time-to-live fields in Ethernet frames.
- As a result of the previous feature, another way to avoid loops in the network is needed.

This lead to the definition of the Spanning Tree Protocol (STP), whose first IEEE standard came through 802.1D-1990 [12]. When STP converges, loops in the network are eliminated, giving as a result a tree whose branches converge on a root node. The root node can be defined manually if the network administrators require it, but there are tie-breaking rules that allow the automatic election of a root node. All paths converge at the root, so all the traffic that goes from one branch to a different one in the network necessarily passes through the root node. STP requires between 30 to 50 seconds to converge properly. As a result, a new definition to improve its speed, the Rapid Spanning Tree Protocol (RSTP) was introduced in 2001 through the 802.1w [13] standard. RSTP is able to respond to a change in the topology in a range of times which goes from a few milliseconds to at most 6 seconds.

With the definition of 802.1 VLANs came the need for defining a separate spanning tree for each VLAN. There were several proprietary solutions from Cisco and Juniper before the IEEE introduced the Multiple Spanning Tree Protocol (MSTP) through the 802.1s standard [14] in 2002. MSTP defines the so called Multiple Spanning Tree (MST) regions. Each MST region can run several MST instances (MSTI). MST regions are interconnected by a Common Spanning Tree (CST). By contrast with proprietary standards prior to MSTP, it uses a Bridge Protocol Data Unit (BPDU) format that has all the MSTIs' information, which reduces protocol traffic inside the network. MSTP is designed to be fully compatible with RSTP. PDUs are built such that, if a RSTP switch-only switch receives a PDU from an MSTP-capable bridge, the MSTI region is seen by the RSTP bridge as a single RSTP bridge.

Nevertheless, all the aforementioned protocols had an important shortcoming: all of them require the definition of a root node which is the meeting point for all the branches of the Spanning Tree. The Spanning Tree is supposed to guarantee a loopless topology, but not the shortest path among bridges. This was solved by the 802.1aq standard [15], that defined Shortest Path Bridging. Shortest Path Bridging uses IS-IS as the link-state algorithm to build the forwarding table of the bridges.

Another approach for shortest path bridging comes from the IETF through the Transparent Interconnection of Lots of Links (TRILL) protocol [16]. Bridges that implement TRILL are known as Routing Bridges (RBridges). TRILL bridges run the IS-IS link state protocol, which is a common feature with 802.1aq. Each RBridge knows about all other RBridges

in the network and the connectivity between them. This allows for calculating shortest path trees to all other RBridges in the network.

All the aforementioned advancements in STP, plus others that will be detailed in the following sections have allowed Ethernet to jump to the MAN and WAN environments. Before this process began, Ethernet lacked scalability to provide several separate VLANs in a backbone environment. Furthermore, it also lacked a Network Management Plane to establish and tear connections. To add more, Ethernet clearly needed QoS and Fault Management, which are essential in a large network. To this end, the 802.1 working group extended the 802.1Q standard to create the 802.1ad, 802.1ah and 802.1Qay standards. In the next sections,we revew the different Ethernet standards that allow Ethernet to move to the Metro.

802.1Q Virtual Local Area Networks

802.1Q virtual LANs is an extension of the Ethernet switching to provide several separated LANs in a network. These LANs might share the same set of bridges, but they cannot see each other from a layer 2 point of view. The reason to have VLANs in a certain organisation is to provide broadcast domain limitation, which reduces congestion, as well as a means to provide security by isolating hosts belonging to a VLAN from other VLANs.

802.1Q uses a 12-bit tag in the middle of an Ethernet frame, called the Q-tag, which is used by bridges to make a logical separation between LANs, giving the users the impression of having several separated LANs in one physical LAN.

The protocol header of an 802.1Q frame can be seen in Fig. 1.6. The Q-tag is inserted between the Ethernet addresses and the Ethertype field. The first 2 bytes of the Q-Tag comprise an EtherType field which specifies that the next two bytes are an 802.1Q VLAN tag. Of those two bytes, only 12 bits are used for VLAN tagging, leading to a maximum of 4096 possible VLANs. In practical terms, this number is much lower because some VLAN codes are reserved.

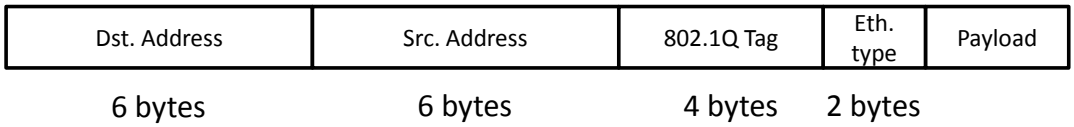


Figure 1.6: 802.1Q header format.

When a frame for another host of a VLAN arrives at one of the switches, it is tagged at the ingress switch and sent through the corresponding trunk port, where the traffic of several VLANs may be mixed. When the frame arrives to the egress switch, it is untagged and sent through the egress port. In order to allow connectivity between VLANs, a router may be connected to each VLAN port, so all stations have a layer 2 connection with the router and the router allows a layer 3 connection between hosts of different VLANs.

In the 802.1Q Ethernet header there are 3 bits which are reserved for defining user priorities, which are mapped to traffic classes. The first definition of traffic classes was given in 802.1p. The 802.1D-1998 [17], to which 802.1p was incorporated, considers 7 traffic classes for different services (voice, video, best effort, etc).

Each traffic class is given a different priority level. Basically, the distinction of user priorities and traffic classes intends provide service differentiation. These definitions were later

included in 802.1Q-1998 [18], without mapping between traffic classes and user priorities (which were included in 802.1D-1998) and the 802.1Q-2005 revision [19], which include the aforementioned mapping. It is worth mentioning that this mapping follows closely the IETF Diffserv paradigm [19, 20].

Unfortunately, 802.1Q is oriented to be used within the boundaries of a single organization, not by a provider. It allows logical separation of VLANs but it is not possible to build more than 4094 VLANs (4096 minus 2 for administration purposes). This number of tags might be sufficient for a corporate network, but not for a MAN. Owing to these reasons, subsequent standards were defined to improve scalability and better adapt Ethernet to MAN and WAN environments.

IEEE 802.1ad Provider Bridges.

To increase the number of supported VLANs, 802.1ad [21] simply adds an additional tag to the C-tag (Customer tag, previously called Q-tag). This tag is called the S-tag and the mechanism is often referred to as Q-in-Q networking. Each service is given a different S-tag but this does not imply any difference in the access to the medium 802.1ad uses the same MAC as classical 802.1Q. As a result, core switches have to learn end station MAC addresses.

The structure of a 802.1ad header can be seen in Fig. 1.7, where the outer 802.1Q Tag is the provider tag, the 802.1Q inner Tag is the customer Tag and the payload field is the customer frame.

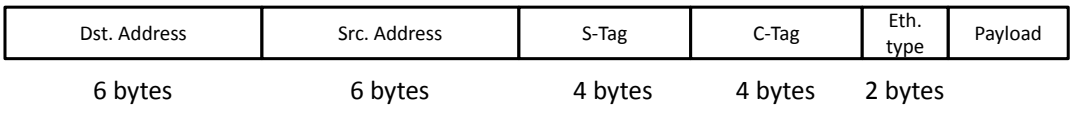


Figure 1.7: 802.1ad header

Each site Outer Tag defines an S-VLAN. Each S-VLAN has its own Spanning Tree, which is calculated by MSTP (Multiple Spanning Tree Protocol). An important issue that has to be taken into account is that the Spanning Tree information of the S-VLANs is known to the Provider Bridges Network.

The main problem of this approach is that core switches have to learn MAC addresses, which implies a great burden for the network to maintain such large forwarding tables. In addition, there is also the fact that S-tags also have 12 bits, and only 4094 services can be defined. Furthermore, changes triggered by STP (Spanning Tree Protocol) can affect the provider network because there is not a formal separation between the customer and the provider's network [6]. Additionally, from the customer's point of view, there is a security risk since addresses are visible outside their domain. That is, there is no differentiation between customer and provider MAC addresses. This lack of separation is also a problem for control protocols. Mainly, the lack of isolation between customer and provider networks, the lack of transparency for control protocols and core switches and the short length of S-tags (only 12 bits) are important drawbacks which prevent 802.1 from having enough (although improved) scalability for a carrier-size scenario. These shortcomings would be solved by the 802.1ah specification.

IEEE 802.1ah Provider Backbone Bridges

The 802.1ah [22] specification adds an additional MAC header intended only for the backbone bridges as well as a Backbone VLAN ID (B-tag, 12 bits) and a backbone service ID (I-Tag or I-SID, 24-bits). This means that up to 2^{24} (around 16 million) services can be discriminated, eliminating scalability problems from previous 802.1 releases in terms of service identification. The 802.1ah header can be seen in Fig. 1.8. B-DA is the Destination Address, that is, the address of the egress switch in the network. B-SA is the Source Address, that is, the address of the ingress switch in the network. B-VID is the Backbone VLAN ID, which will be explained later. I-SID is the Service Instance Identifier, which will also be explained later. The rest of the frame might be a plain Ethernet frame, an 802.1Q frame or an 802.1ad frame, depending on the customer and the service provider's network organisation. It is worth noting that it is not compulsory to have an 802.1ad frame inside a 802.1ah frame. Customer frames can be inserted directly into the MAC-in-MAC level.

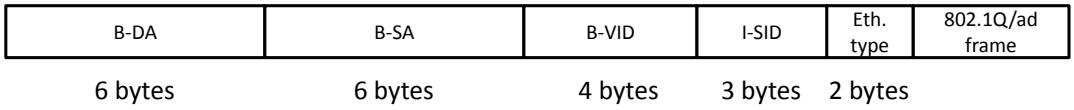


Figure 1.8: 802.1ah header.

From the architectural point of view, PBNs and PBBNs operate in the following way [23]. A single PBBN is composed of two different bridge types:

- Backbone Edge Bridges (BEB). These bridges are in the outermost parts of the network and their duty is to encapsulate and decapsulate the frames from and to the PBNs or customer networks directly attached to them.
- Backbone Core Bridges (BCB). These bridges transport frames within the core of the network. Frames are carried on provisioned Backbone Service Instances, represented by BSIs (I-Tag), which are carried on B-VLANs (B-VIDs). The BSIs can be point to point, point to multipoint or multipoint to multipoint. PBBNs support several kinds of attachment to the network:
  - Direct attachment from a PBN.
  - Attach to a network by means of a port-based interface or an 802.1ad adaptation layer.

An example of this architecture can be seen in Fig. 1.9. Each level of the hierarchy supports up to 16.776.059 service instances. This hierarchy might be extended by joining PBBNs [22, 23]. 802.1ah provides two kinds of interfaces for multidomain PBBNs, as can be seen in Fig. 1.10.

1. Hierarchical interface: In this case, frames from the lower PBBNs are encapsulated into frames of the higher PBBNs. This kind of encapsulation may be performed until the maximum frame size is achieved. Each new level treats the lower one as a PBN, multiplying the number of service instances by 16.776.059 [22].

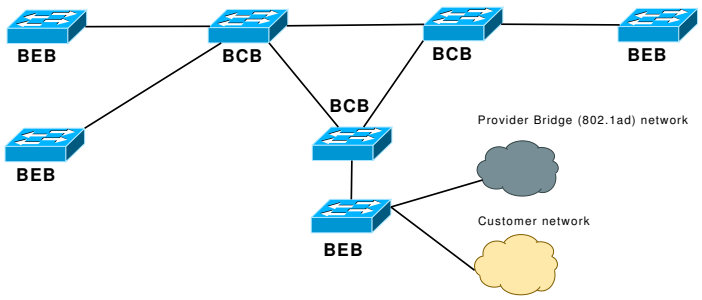


Figure 1.9: 802.1ah inner architecture.

2. Peer interface: There is no encapsulation in this hierarchy. In peer networks, the number of service instances can duplicate the number of service instances in one hierarchy in the best case. The maximum number of service instances would be 33.553.918 (supposing that there are no service instances between the nodes of the peer networks). This number would decrease as service instances having end points in different peer networks appear [23].

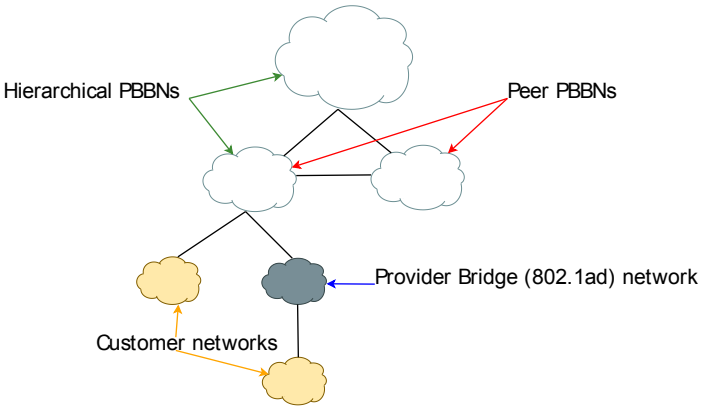


Figure 1.10: 802.1ah architecture relations.

The I-SID is used for service identification and the B-Tag can be used to segregate the network into regions with different capabilities and technologies on each one. An example could be:

- Use some Backbone VLAN IDs to provide E-Line services with traffic engineering.
- Use some other Backbone VLAN IDs to provide E-LAN and E-TREE services without traffic engineering.

A B-VID can also be used to support several E-Line Services between points of presence. Individual services can then be configured in the edges of the network, while in Provider Bridges the services must be configured node by node, which implies an important operational burden.

The forwarding mechanism is based on the Backbone Destination Address (B-DA) and the Backbone VLAN ID (B-VID or B-Tag). The I-Tag identifies Service Instances at the edges of the network, but it is not used for forwarding in the core network. For point to point services, an example is shown in Fig. 1.11. When a frame from a customer or provider network arrives at a Backbone Edge Bridge (BEB) and this frame is for a node which is in another site, the customer frame is tagged with an I-Tag (I-Tag A) to identify the Service Instance that connects both sites of Customer/Provider A. The frame enters the network and it is forwarded according to its B-VID and its B-DA. When the frame reaches its destination, it is decapsulated and associated to Customer/Provider A due to I-Tag A. This implies that BEBs have to learn the addresses of end stations and map them to their respective Service Instances as well as to other BEBs holding the same Service Instance. That is, BEBs have to learn both end station and BEB MAC addresses. By contrast, BCBs need only to learn the addresses of nodes of the core network to forward frames.

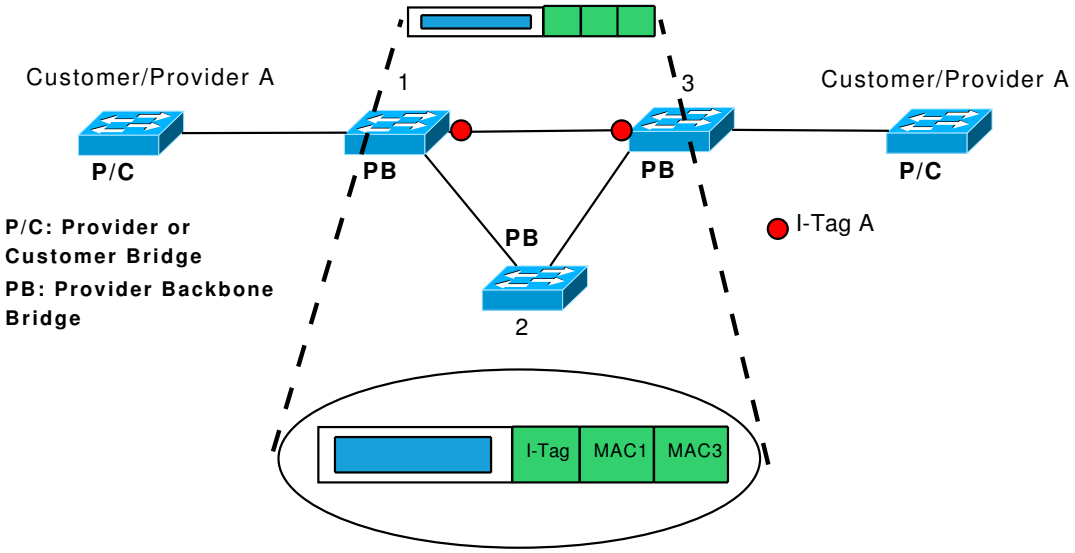


Figure 1.11: Forwarding process in a Point To Point Service.

For multipoint services, it depends on whether it is a unicast or broadcast frame. For unicast frames, the mechanism is shown in Fig. 1.12. When a frame arrives to a Backbone Edge Bridge (BEB), it looks in its switching table for an entry to the destination MAC address of the Customer/Provider frame. The bridge finds the associated I-Tag to the instance service, the B-VLAN ID and the B-DA where the customer station is. The frame is sent to the B-DA of the BEB holding the station [22].

For broadcast frames it is slightly different. Each I-Tag is associated with a B-VID and each Backbone VLAN has its own spanning tree. Broadcast frames from one customer are sent in broadcast mode through the tree of the B-VLAN to which the I-Tag is associated. This is shown in Fig. 1.13. In this way, broadcast frames get to all the BEBs that have that I-Tag configured. The first 24 bits of the B-DA of these frames are the prefix 00:1E:83 and the last 24 bits are the I-SID of the service instance [22].

However, this forwarding mechanism is extremely inneficient since the broadcast frame gets to all BEBs and it is filtered at the destination BEB by means of the I-Tag. There are

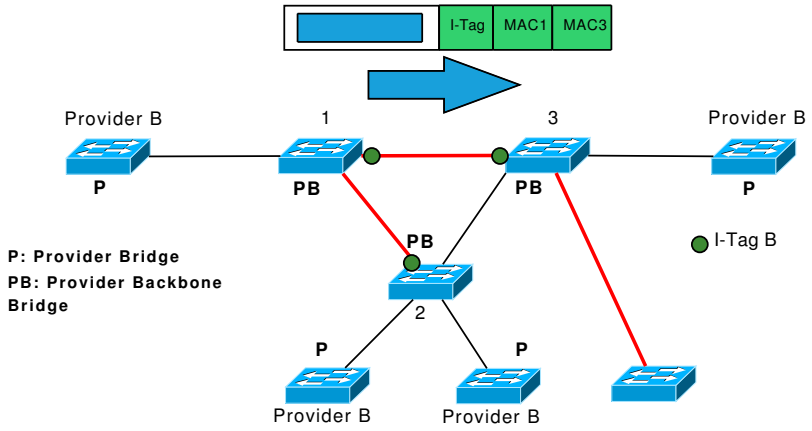


Figure 1.12: Forwarding process in a Point To Multipoint Service (Unicast case).

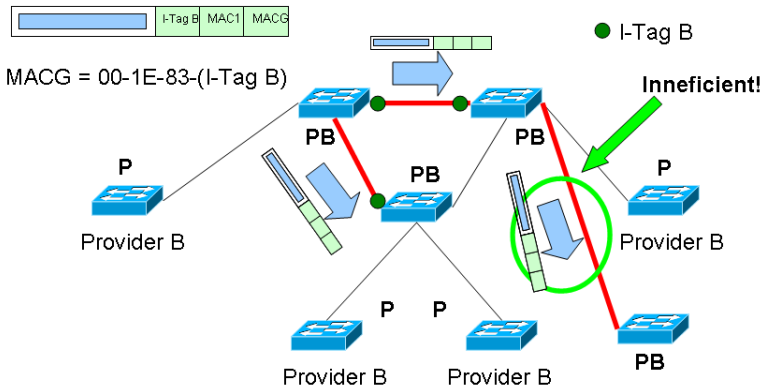


Figure 1.13: Forwarding process in a Point To Multipoint Service (Broadcast case).

some BEBs that do not have that I-Tag and are receiving frames innecessarily. A way to prevent innecessary delivery of frames would be to apply delivery lists so Backbone Bridges are able to know which frames are to be delivered to which BEBs [22]. This procedure saves bandwidth and is shown in Fig. 1.14.

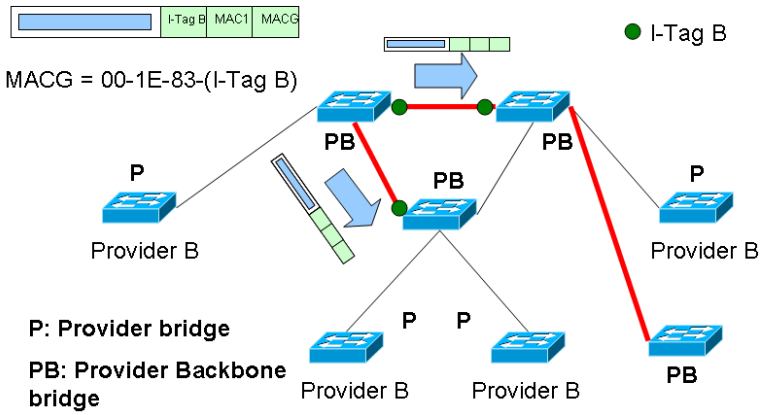


Figure 1.14: Forwarding process in a Point To Multipoint Service for broadcast frames with delivery lists.

Because 802.1ah defines an additional MAC header, scalability is significantly improved because the switches in the core network need only see the outer MAC header to perform the forwarding function, ignoring the inner customer headers eliminating the need to learn customers' addresses in the forwarding table. Another indirect advantage of this is that, as the forwarding is only based on the outer MAC header, only the switches in the edge of the backbone network need to be Provider Backbone Bridges. Switches in the core network could be provider bridges, as it is only needed that they can forward the frames according to the destination MAC address. In this case, the S-Tag would have the same Ethertype as the PBB B-Tag and the forwarding would be based on the B-DA and B-Tag. Provider Bridges can assign C-VIDs and S-VIDs knowing that PBBs will not change them. Furthermore, all operational Ethernet protocols are tunnelled transparently across the network. In this way, the customer can use these protocols without affecting the provider's network. In other words, with 802.1ah improved scalability is achieved.

Improved scalability is also attained thanks to the I-SIDs (24 bit long), which allow us to define a great number of service instances, eliminating the problems of 802.1Q and 802.1ad. B-VIDs allow the partitioning of the network and dividing it into regions with different capabilities. Because all of this, 802.1ah represents a great step into converting Ethernet in a carrier technology but there are still certain issues to be solved. Mainly, these issues are Traffic Engineering and OAM (Operation and Management), which are addressed in section 1.3.1.

IEEE 802.1Qay Provider Backbone Bridges-Traffic Engineering.

In the previous sections, the way how service instances are identified is defined. When a frame from one customer enters the network, it is mapped to a service instance identifier

(I-SID) and forwarded through the backbone towards the BEBs supporting that service instance. The 802.1ah specification does not define how paths through destinations are discovered. That is, how one BEB knows which other BEBs support a traffic instance. Because of this, BEBs have to flood the whole B-VLAN associated with the I-Tag to deliver broadcast frames or frames for unknown stations. The IEEE 802.1Qay, Provider Backbone Bridges Traffic Engineering (PBB-TE) [24] standard allows to define end-to-end Ethernet paths by configuring hop by hop the switching tables of Backbone bridges.

To better explain this, let us define serveral key concepts about PBB-TE:

1. Provider Backbone Bridge Network (PBBN). A network that uses Backbone Core Bridges (BCB) and Backbone Edge Bridges (BEB), which are defined in the IEEE 802.1ah standard.

2. Backbone Service Instance. It is an instance of the MAC service in a PBBN (I-Tag) between two BEBs.

3. PBB-TE Region. A contiguous set of BEBs and BCBs capable of providing TE service instances. Bear in mind that the standard does not specify that a PBB-TE region should include all the Backbone Bridges in the network.

4. TE service instance (TESI). An instance of the MAC service supported by a set of Ethernet Switched Paths (ESPs). Each TESI is identified by TE-SID, forming a bidirectional service. TESIs can be point-to-point or point-to-multipoint. Each TESI is identified by a series of three tuples which identify the ESPs associated to the instance.

5. ESP (Ethernet Switched Path). An end-to-end path in a PBB-TE Region. The path is configured by adding static entries in the FDB of the bridges that conform the path. Each ESP is uniquely identified by a three-tuple: (ESP-MAC DA, ESP-MAC SA, ESP-VID). ESP-MAC SA is the MAC address of the BEB where the ESP starts, ESP-MAC DA is the MAC address of the BEB where the ESP ends and ESP-VID is a VLAN ID used in a special way which will be explained later.

6. External Agent. It is not formally defined in the standard. The role of the External Agent is to configure the Filtering Databases (FDBs), that is the Switching Tables, of Backbone Bridges. When an ESP is to be established in the network, the External Agent adds the appropriate entries in the Switching Tables of all the bridges that conform the ESP. The architecture of and protocols used by the External Agent are not specified in the standard. The External Agent acts within the scope of a PBB-TE region. The External Agent is allocated a range of ESP-VIDs to establish ESPs.

One or more Service Instances, which are identified by an I-Tag, can be associated to an ESP [24]. When a frame from a Service Instance is to be forwarded through the network, it is introduced in the ESP. Then, the frame is forwarded towards the egress of the tunnel. The frame is associated in the egress to the service instance by the I-Tag.

ESPs are, in practical terms, Ethernet unidirectional tunnels. Each ESP is uniquely identified by the 3-tuple (ESP-MAC DA, ESP-MAC SA, ESP-VID). Only the duple <ESP-MAC DA, ESP-VID> is used for forwarding in the core network [24]. This duple is configured

statically by the External Agent in each of the DFBs of the bridges that belong to the switching path. It is treated as a unique 60-bit ID (<48,12>) by the backbone bridges.

ESP-VIDs are B-VIDs which have been marked to be out of the management of the spanning tree protocols [24]. ESP-VIDs are managed by the External Agent or Control Plane. A capital difference between 802.1Qay and the rest of Ethernet specifications is that ESP-VIDs are treated locally rather than globally in comparison with B-VIDs.

Traditionally, B-VIDs are VLAN Tags which simply identify a VLAN. The Multiple Spanning Tree Protocol (MSTP) runs over each VLAN instance. After the protocol converges, a Spanning Tree associated to each VLAN instance is created. When a frame with a certain VLAN Tag enters a bridge, it is forwarded through the links belonging to the Spanning tree of that VLAN. Thus, VLAN tags have a global meaning. In the example, there are two VLANs in the Backbone Network, each with a different VLAN ID. Should the same VLAN ID be used for two different VLANs, it would be impossible for a bridge to distinguish which frame belongs to each VLAN. This is very different in PBB-TE, where the 3-tuple (ESP-MAC DA, ESP-MAC SA, ESP-VID) identifies uniquely each ESP.

There are two types of TESIs, the point to point TESIs and the point to multipoint TESIs. Point to Point TESIs are shown in Fig. 1.15. It is formed by two co-routed ESPs, that is, two unidirectional ESPs that follow the same path in the network and define in this way a bidirectional service [23, 24]. The point to multipoint TESIs are shown in Fig. 1.16. They are formed by one multipoint tree ESP that has n leaves (endpoints) and n point to point ESPs from the leaves to the source [23, 24]. The point to point ESPs from the leaves to the root are co-routed along the branches of the tree. The ESP-DA of the point to multipoint is a group MAC address that identifies the n leaves of the tree.

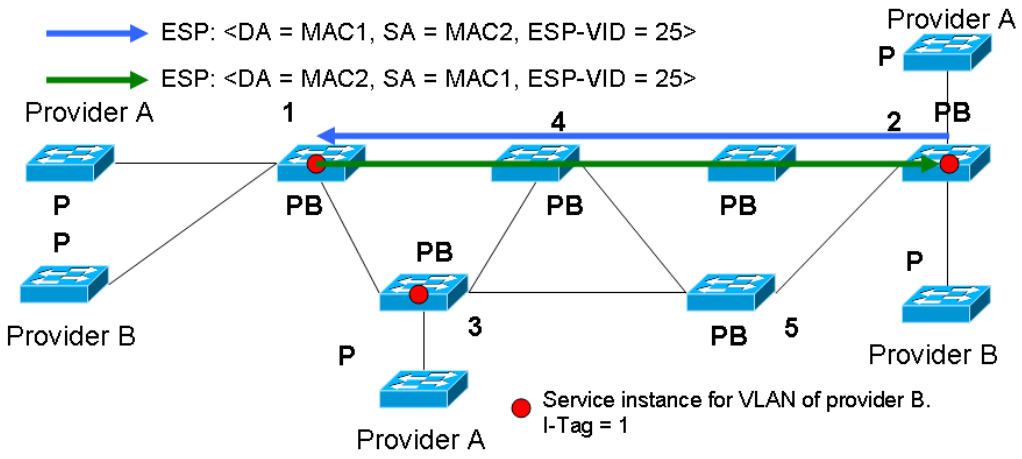


Figure 1.15: A point to point TESI.

Connectivity Fault Management and performance monitoring

Further extensions have been made to Ethernet to ensure that if a failure occurs, it will be reported to the operator. Connectivity Fault Management (CFM) is a capital feature for operators, as it is important to diagnose and solve network failures as quickly as possible.

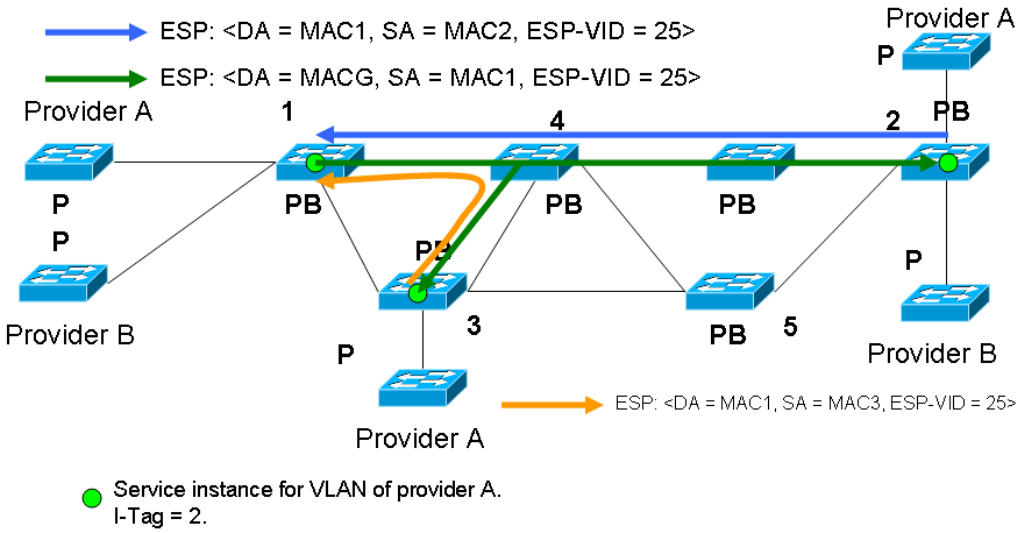


Figure 1.16: A point to multipoint TESI.

There are four mechanisms for fault management in Metro Ethernet [6, 24, 25]. The four mechanisms are:

- **Fault detection:** Continuity Check Messages (CCMs) are sent very 10 ms (this period is configurable by the network manager). If three consecutive CCMs are lost, a failure is considered to have happened. Then, an alarm is issued to the network management plane.
- **Fault notification:** According to the ITU-T Y.1731 standard, devices can be configured to report Alarm Indication Signals (AISs) to the network management plane if a failure is suspected to have occured. The triggering event for this situation could be either three lost CCMs or some other event that might give a clue about a failure. The failure should be found and isolated by the two mechanisms explained below.
- **Fault verification:** If failure is suspected to have ocurred, a lopback message (LBM) is sent to a specific destination. The destination will reply with a loopback reply (LBR). Of course, if a failure has effectively taken place, there will be no LBR messages for response. The main difference between CCMs and LBM/LBR messages is that CCMs are issued periodically and LBM/LBR messages are sent manually by the network operator to locate the failure.
- **Fault isolation:** Another pair of messages is defined for fault isolation. One node in the network issues a linktrace message (LTM) towards an end node. Each intermediate node has to reply with a linktrace reply (LTR) to the source of the LTM message. This allows failure isolation by seeing which is the first hop that does not reply to the linktrace message.

It could be said that the fault verification and fault isolation mechanisms have certain parallelism with the ICMP echo messages to verify the reachability of a node and perform

traceroute operations, with the key difference that fault verification and fault isolation are defined for the Ethernet layer. Of no lesser importance for an operator is to be able to monitor the network performance. The ITU-T Y.1731 standard provides the functionality to monitor basic performance metrics [6] such as:

1. **Frame loss ratio:** this metric is calculated by means of the ETH-LM (Loss Measurement) counters in the CCM messages to know the number of lost frames and thus calculate this ratio.

2. **Frame delay:** this requires endpoints to have synchronized clocks. The delay is calculated by means of the ETH-DM (Delay Variation Information) counters in the CCM messages.

3. **Frame delay variation:** this metric is calculated by using frame delay measurements.

The metrics mentioned above are tracked and certain alarms can be issued to the network operator if certain thresholds are exceeded. This is especially useful for Service Level Agreement (SLA) enforcement.

Energy Efficiency in Ethernet

Other extensions of the Ethernet technologies are related to the growing need of energy efficiency in communications. Energy efficiency has been the subject of research for over a decade [26, 27]. The ever increasing presence of Information Technologies and the need for electronic communications has significantly highlighted the importance of reducing energy consumption. After all these years, there have been efforts in many fields to reduce this energy consumption, giving the scientific community a perspective of the areas in which to improve efficiency [28]. There are four main lines of research regarding energy efficiency, as pointed out by [28].

The first main line is that of *adaptive link rate*. This family of techniques involve two main subfamilies, *sleeping mode* and *rate switch*. *Sleeping mode* techniques [26, 29, 30] consider only two states of operation which are sleeping (idle) or full activity and they try to find a compromise between energy savings and the adaptation to idle periods. *Rate switch* techniques involve, as their own name indicate, adapting network link rates to operation conditions (for instance, traffic load) [31–33]. Both of the aforementioned approaches have been implemented into the Energy Efficient Ethernet (EEE) standard, but only the first (known as Low Power Idle (LPI) has survived the standardisation process, while the other one, Rapid PHY Selection (RPS) was later dismissed.

The second line is *Interface Proxying*. *Interface proxying* in delegating the processing of network packets to other entities. *Interface proxying* can be further divided into *NIC proxying* [31, 34, 35], if the processing is delegated to the integrated NIC of the network node, and *external proxying* [31, 34, 36], if the processing is delegated to other nodes in the network.

The third line of research in energy efficiency is *Energy aware infrastructures* [37, 38] aims, among othe things, on the role of network design in energy consumption. Finally, the fourth line, *Energy aware applicatons* involves changing the way in which the application level works by suggesting energy efficient application [39, 40] and transport [41] protocols.

Of all the lines indicated previously, one of those which has got a great deal of attention is Energy Efficient Ethernet (EEE) [3]. Energy Efficient Ethernet tries to achieve energy savings by sending the Ethernet Interface to a low power state when there are no frames to transmit in the link. This state is known as Low Power Idle (LPI). A sample of operation can be seen in Fig. 1.17. In Energy Efficient Ethernet, a set of times, namely T_w (time to wake up the interface from LPI), T_s (Time to transition the interface to the LPI state) and T_r (refresh time) are defined. When an interface has not packets to transmit, it goes to sleep, which takes T_s seconds. If a frame arrives, then it takes T_w seconds to wake up, transmit the buffered packets and then go to sleep again. In Fig. 1.17, a single frame transmission is displayed and it T_{Frame} represents the frame transmission time. To not lose synchronisation completely, with a certain regularity synchronisation frames are transmited, which takes a time T_r.

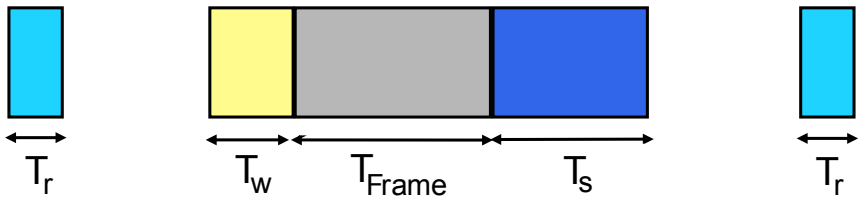


Figure 1.17: Example of EEE operation

The efficiency of EEE can be defined according to Eq. 1.1, considering T_r despicable:

$$Eff = \frac{T_{tx}}{T_w + T_{tx} + T_s} \tag{1.1}$$

Where T_{tx} is the transmission time of all the frames buffered before the interface wakes up. A table with the minimum times for the timers mentioned, as well as their associated efficiencies, is given in Table 1.1. It can be observed that, for low values of T_{tx}, the efficiency is poor, which means that short packets put a significant penalty on efficiency. This efficiency can be so low as 0.6% in the case of 150B frames in 1000Base-T.

Protocol	Min T_w (μs)	Min T_s (μs)	T_{Frame} (1500B) (μs)	Frame eff.	T_{Frame} (150B) (μs)	Frame eff.
100Base-Tx	30	100	120	48%	12	8.5%
1000Base-T	16	182	12	5.7%	1.2	0.6%
10GBase-T	4.16	2.88	1.2	14.6%	0.12	1.7%

Table 1.1: Minimum values for the timers specified in the IEEE 802.3az standard [3] and the efficiency values for two frame sizes in bytes, taken from that of [4]

While achieving significant energy savings, certain performance issues have been addressed [4], especially in bursty environments. This is due to EEE not requiring a minimum number of packets to transmit each time the interface is waken up.

Further energy savings with the coalescing of packets can be achieved [42,43], since that way it is possible to prevent waking up the interface (which has an associated energy cost)

to transmit a small number of packets. Let us consider a simple example to illustrate this. According to Table 1.1, a 150B frame in 1000Base-T can have a efficiency so low as 0.6%. If instead of transmitting a single 150B frame we would transmit 2000 in a row, applying Eq. 1.1 we would obtain:

$$Eff = \frac{T_{tx}}{T_w + T_{tx} + T_s} = \frac{2000 \cdot 1.2}{16 + 2000 \cdot 1.2 + 182} = 0.924\ (92.4\%) \tag{1.2}$$

This simple example shows that, with coalescing, proper savings can be achieved. Of course, there is a trade off between packet coalescing and delay that has to be taken into account.

1.3.2 Optical transparent networks based on WDM ring architectures

A Wavelength Division Multiplex (WDM) optical network is an optical network in which several channels or wavelengths can coexist, thus multiplying transmission capacity. Each channel transports independent information from the other channels multiplexed in the same time frame. However, WDM is not that much a novelty as a transparent WDM network. With the increasing bitrate in optical networks the switching speed at the electronic level has risen accordingly. The constraints of the electronic medium have thus become more and more apparent. It is for this that the research in optical transparent networks has taken the focus in the last decades. Transparency in an optical network implies all-optical transmission, that is, without OE and EO conversion. In a WDM transparent optical network there are several wavelengths circulating at the same time.

The progressive orientation of Ethernet to the MAN and WAN seen in previous sections is no coincidence and goes parallel with the advancements in the field of optical networks. It is because of this that new MAN and WAN architectures have taken the biggest part of the focus in the transparent optical network research. One of the most renowned architectures for MAN and WAN networks are ring networks due to their simple and effective topology to cover a certain area, especially since it is the preferred architecture in legacy SDH networks. In this sense, transparent WDM ring networks seem to be the logical next step.

Transparent WDM ring architectures can be classified according to the architecture and the Medium Access Control (MAC) scheme used. One aspect to take into account is the laser/photodiode scheme used in each node. Because there are multiple wavelengths in the network, a WDM ring might have multiple transmitters, multiple receivers or both. Because lasers can be expensive, some architectures might have either tunable transmitters or tunable receivers or both to reduce costs.

If a laser is fixed, that is, that is, it has a single wavelength, we call it a *Fixed Transmitter* (FT). The same can be said respectively for those photodiodes which are tuned to a single wavelength, which are known as *Fixed Receivers* (FR). Similarly, we define *Tunable Transmitters* (TT) and *Tunable Receivers* (TR) for tunable lasers and tunable receivers. A WDM ring network might have zero, one or several of the aforementioned transmitter or receiver pairs and we denote them indicating the number of each kind in a superscript. For instance, a network in which each node has four fixed transmitters, two fixed receivers and a pair of tunable transmitter and fixed receiver would be denoted as $FT^4 - FR^2\ TT - FR$.

Another aspect to classify WDM transparent ring architectures comes from how do they access the medium both in the passive (listening) and active (transmitting) aspect. In the

passive sense, some architectures use channel inspection to avoid collisions, others do not. In the active sense, some architectures divide the channel space in slots of fixed size, while others allow more flexible schemes. In the next sections, some well known WDM ring architectures and laboratory testbeds are presented. These are MAWSON, DBORN, RINGO and HORNET.

- MAWSON: The Metropolitan Area Wavelength Switched Optical Network (MAWSON) [44] is based on a $FT^W - FR$ or $TT - FR$ scheme. There are N nodes connected to the ring by means of passive Optical Add-Drop Multiplexers built upon Fibre Bragg Gratings (FBGs) to drop specific wavelengths for reception at a node. Each node has a dedicated home wavelength. In an architecture like this, multicasting and broadcasting could be achieved by turning on several lasers on the $FT^W - FR$ variant.

  MAWSON is a slotted ring in which slots are aligned for all wavelengths. In this WDM ring architecture, there is no channel inspection but rather a request-response protocol among nodes. If node i wants to send a packet to node j, node i issues a request to node j, who then allocates certain one or several on its home wavelength for node i, which then transmits the desired packet.

- RINGO: The Ring Optical Network (RINGO) [45] is a slotted ring with channel inspection with an $FT^W - FR$ architecture, having one wavelength per receiving node. In this network architecture, home wavelength inspection is implemented to sense the channel and Virtual Output Queueing (VOQ) is used to prevent head of the line blocking. This means that, when a node detects an empty slot in any of the wavelengths it wants to transmit in, it inserts the packet it wants to transmit in the network. A main drawback of this architecture is that, since there is no reservation on slots, it is not possible to control how many slots can be taken by a node and that leads to a need for fragmentation in packets, since slots are fixed in size.

- HORNET: The Hybrid Optoelectronic Ring Network (HORNET) [46] has a $TT-FR$ scheme and, as the other previous architectures studied, it uses destination stripping. It also uses VOQ and channel inspection as RINGO does, but instead of decoding each wavelength separately to see which slots are free in each possible wavelength, it can implement two sensing schemes, which are outline in [46]. The first one was a Carrier Sense Multiple Access/Collision Avoidance (CSMA/CA) based scheme in which each wavelength had an associated sub-carrier which indicated if the specific slot was occupied, but it was later dropped in favour of a network architecture with a dedicated control channel for all wavelengths. The control wavelength is dropped at every node and a chain of bits in the channel indicate which wavelengths of the next slot are free, so the node knows in advance the available transmission slots.

- DBORN: The Dual Bus Optical Ring Network (DBORN) [47] is a WDM ring network in which a hub node receives all the upstream traffic from the nodes and it forwards it all through the downstream channels to the reciver nodes. The ring is transparent from the nodes to the hub node, but optical to electrical conversion is performed at the hub node. If two nodes need to send information to each other, they must send it

first to the hub node. The hub node also acts as a gateway to other networks. DBORN relies on a CSMA-like solution, so it does not need fixed time slotting. There are a set of upstream wavelengths and a set of downstream wavelengths. The nodes sense the upstream wavelengths with low bitrate photodiodes and, when inactivity is detected on the link, packets are inserted on the upstream wavelengths if the period of inactivity is long enough for the packet to fit in it. In the downlink, several nodes can share a certain wavelength and they have to inspect which packets are directed towards them. That is, in the upstream direction there is a share in writing and in the donwlink direction there is a share in reading.

A simple example on a TT-FR architecture

Having seen the main architectures and testbeds for optical WDM ring networks, it is time to see a simple example that help the reader visualise how a TT-FR WDM ring might work in terms of access to the medium. In Tunable-Transmitter Fixed-Receiver (TT-FR) WDM ring topologies with N nodes and W wavelengths, each node in the ring has got a tunable laser (tunable to any wavelength) and a receiver, which is always fixed to the same reception wavelength. In the cases where $N = W$, each node in the ring can be provided with its own reception wavelength (often called *dedicated home channel*) and, at the same time, the tunable laser provides each node with a means to communicate (directly) with all others, just by tuning its laser to the appropriate reception wavelength of the destination node. Transparency is achieved because all nodes by-pass all wavelengths except its dedicated home channel, which is stripped off the ring by its own node. Thanks to WDM and transparency, each node is provided with a logical unidirectional channel to all other nodes in terms of transmission, but it can only receive frames from a single wavelength [44].

A simple conceptual example can be seen in Fig. 1.18. In our example, there are four nodes numbered from 1, to 4, each of them with a home wavelength (red, green, blue and purple). In the example, node 3 wishes to send a frame to node 1 and node 4 wants to send a frame to node 2 at the same time. Node 4 tunes its tunable laser to node 2's wavelength and sends it transparently through the network (there is no optical to electronical conversion at node 3). Node 2 has a fixed receiver, so it just receives the frame directed to it. The same process happens between nodes 3 and 1. Because the ring is WDM, there is no collision between both transmissions as they travel in different WDM channels. Medium Access Control as indicated in the previous techniques would be required if two nodes were to transmit a frame to the same destination node.

The TT-FR architecture has been widely studied in the literature because of its simplicity, and a number of MAC protocols have been defined in order to arbitrate channel access, such as: RingO, Mawson, DBorn [44].

Concerning the broadcasting of frames, the easiest implementation considers the replication of the frame over all destination home channels. However, this strategy is clearly bandwidth consuming, especially when the number of nodes in the ring is large.

Alternatively, broadcasting a frame around a ring can be implemented by sending a single frame hop-by-hop along the ring following a store-and-forward basis. This strategy scales with the number of nodes in terms of bandwidth consumption, but has two main drawbacks: (1) the frames must suffer OEO on each node, thus increasing latency (and removing

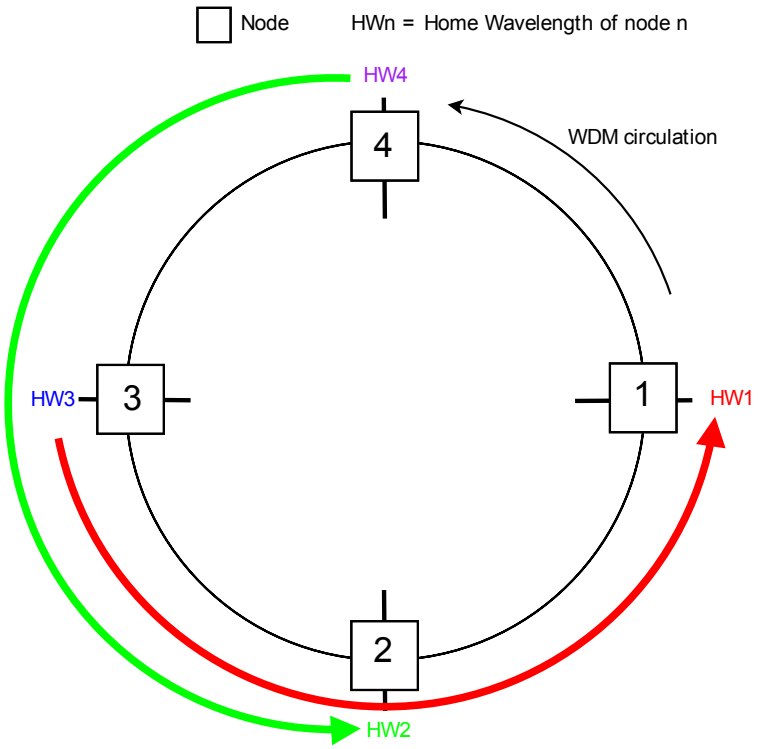


Figure 1.18: A simple example of a TT-FR WDM ring.

transparency) and (2) some extra control is necessary: the source node must remove its frame after a loop around the ring.

1.3.3 Optical technologies based in OCDMA

Introduction to OCDMA

Code Division Multiple Access (CDMA) has been used for many years in radio systems, being 3G mobile telephony the most well-known example of this. It is probably because of this and the ever increasing need of efficiency in the spectrum use for optical transmission that made the adoption of OCDMA in the optical domain one of the many alternatives considered along with WDMA, which was the object of study of Chapter 2.

CDMA is a technique which arbitrates the access to the medium based on the use of orthogonal codes. By contrast with Time Division Multiple Access (TDMA), which regulates the access to the medium through the use of differentiated timeslots and with Frequency Division Multiple Acess (FDMA), which uses different frequency slots, CDMA allows several users to access the same spectrum at the same time. A CDMA system is defined by a CDMA code space. Each code in a CDMA code space allows one user to access at the same time and spectrum without (ideally) interfering with any other user. This is made by implementing orthogonal (or quasi-orthogonal codes). With an appropriate decoder, the receiving end is able to discriminate the transmitted signal from the signals of the rest of the users. In order to achieve this, CDMA codes have the following important properties [48]:

- A codeword has high autocorrelation, having a distinct peak and low autocorrelation sidelobes. This property allows that the receiver end discriminates one specific user from the others.
- Two any different codewords of the code space have a very low (ideally zero) cross correlation. This property allows that the receiver end does not receive too much interference from other users' signals
- One codeword must be distinguishable from a shifted version of itself.
- One codeword must be distinguishable from a shifted version of any other codeword. This property and the previous one allow proper synchronisation between the receiver and the transmitter.

Nevertheless, while the natural use of CDMA in the radio medium has been mostly based on (ideally) orthogonal codes, this does not apply as widely in its optical domain translation. In this sense, Optical Division Multiple Acess (OCDMA) is divided into two large groups regarding how codewords are discriminated:

- Incoherent systems: these OCDMA systems base code discrimination in techniques that do not use the vectorial nature of the electricmagnetic field of light. This results in quasi-orthogonality of codes.
- Coherent systems: these OCDMA systems base code discrimination in techniques that use the vectorial nature of the electricmagnetic field of light. This results (almost) orthogonal codes.

Both incoherent and coherent systems suffer a limitation that is known as Multiple Access Interference (MAI), even though coherent systems are supposed to provide ideally orthogonal codes. MAI might be caused by either the non-orthogonal nature of the codes, as it is the example of the incoherent systems or due to the non-perfect nature of coherent systems. MAI increases with the number of nodes accessing the medium and has a direct influence in the Binary Error Rate (BER) of OCDMA systems. When the MAI is so large that the BER reaches unacceptable levels, this is defined as an *outage*. Because this is a gradual deterioration of the signal, OCDMA systems are said to have what is called *soft capacity*, that is, a number of users effectively interfering each other do not cause immediate error in the communication as it would happen in an TDMA or WDMA system, but there exists a threshold above which communication becomes impossible due to excessive interference among OCDMA nodes.

In the following sections, the two main families of OCDMA techniques are presented and the main representative techniques of each family are explained.

Incoherent OCDMA

This family of OCDMA techniques is characterised by not using the vectorial nature of light for modulation and detection. The main representative techniques of this family are:

- Direct-sequence or temporal encoding OCDMA systems.

- Spectral amplitude encoding OCDMA systems.
- Two-dimensional spatial encoding OCDMA systems, also known as spread-space encoding

One of the main techniques in OCDMA is the direct-sequence or temporal encoding OCDMA. This technique of OCDMA encoding is made by coding each bit into a set of chips. If T_b is the bit period and T_c the chip period, such that $T_c < T_b$. Each '1' bit, represented by the presence of a constant pulse of length T_b, is coded as a set of L_c chips of length T_c. Given that there is no sign information since only the light intensity (light or no light) is used as a way of coding chips, all the codes in this family are quasi-orthogonal. That is, they are not perfectly orthogonal.

Said family includes Optical Orthogonal Codes (OOC) and algebraic congruence codes [48]. Algebraic Congruence Codes comprise Prime Codes (PC), Quadratic Congruence Codes (QCC), Cubic Congruence Codes (CCC) and Hyperbolic Congruence Codes (HCC).

OOCs [49–52] are built with the certain properties of autocorrelation and cross-correlation in mind as the main set of restrictions for code design. With this family of codes, better cardinality, autocorrelation and cross-correlation properties are achieved, but its construction is more complex than that of Algebraic Congruence Codes.

Algebraic Congruence Codes are constructed by means of congruence techniques, which make use of what is known as congruence operations finite fields to create the optical codes [48]. This family of codes have a greater regularity and are way simpler to construct than their OOC counterparts, but they achieve sensibly worse cardinality, autocorrelation and cross-correlation properties.

Prime Codes [53], Quadratic [54, 55], Hyperbolic [56, 57] and Cubic [58] Congruence Codes are built by choosing a prime number p and using its associated congruence operation over the corresponding Galois Field $GF(p)$.

Another important incoherent OCDMA technique is Spectral Amplitude Encoding (SAE). SAE divides a broad spectrum in pieces and codes the information in the amplitude of that spectrum [48]. An example implemented with optical bulk components was presented in [59, 60].

The a bulk SAE encoder is composed of a pair of uniform diffraction gratings, a pair of lenses and a spectral amplitude mask. The broadband source provides a broadband beam that is modulated by data. The broadband beam is then decomposed and separated into several distinct spectral components $\lambda_1 \cdots \lambda_n$ by the first diffraction grating. The first lens improves the spectral separation among components before the light beam going through the spectral amplitude mask. The spectral amplitude mask is an optical filter which selects which components of the broadband spectrum are used to encode the user's information.

After going through the spectral amplitude mask, the rest of the spectral components are combined by the second lens and the second diffraction grating into a single optical beam and is transmitted through the optical network. The composed transfer function of the encoder is denoted as $A(\omega)$. Regarding the receiver end, the signal goes through a decoder composed of a 3dB coupler, two spectral amplitude amplifiers and two differential fotodetectors. The transfer function of the two spectral amplifiers is $A(\omega)$ and $\overline{A}(\omega)$, where $\overline{A}(\omega)$

is the complementary function of $A(\omega)$. This allows differential detection is the cancellation of multiple user interference. The minimisation of multiuser interference depends, of course, on the code design too. SAE OCDMA can also be achieved by using FBGs (Fibre Bragg Grating) [61]. The encoder works by the same principle as that of [59], but diffraction gratings are substituted by FBGs.

Finally, two dimensional approaches are of varied nature and may involve combining several one-dimensional techniques. Nevertheless, specific aspects of these techniques are out of the scope of these thesis. The reader is encouraged to further study such techniques in [48].

Coherent OCDMA systems

Coherent OCDMA systems are characterised by the use of the phase of the optical signal field. This requires the use of a highly coherent source, such as a mode-locked laser [48]. The receivers for these systems rely on a coherent reconstruction of the original optical signal to get the encoded user data.

There are two main coherent OCDMA techniques. One is the Spectral Phase Encoding OCDMA (SPE OCDMA) and the other is the Temporal Phase Encoding OCDMA (TPE OCDMA). While SPE OCDMA encodes the user information in the phase of the optical signal in the spectral domain, TPE OCDMA encodes the user data in the phase of the optical signal in the time domain. SPE and TPE OCDMA are introduced in the following subsections.

The first of the two main techniques studied is Spectral Phase Encoding (SPE) OCDMA. A coherent OCDMA spectral phase encoder model was first explained in [62]. It uses two diffraction gratings located at the focal planes of a unit magnification, confocal lens pair as well as a multi-element phase modulator. During each bit "1" of duration T_b, an ultrashort coherent optical pulse of duration T_c is input into the optical encoder. Then, the first grating decomposes the spectrum in a set of spectral components of the aforementioned pulse. The decomposed lightbeam goes into a spatial phase mask according to the user codeword. The codewords follow a bipolar code.

The mask introduces phase shifts among the spectral components following a pseudorandom previously established pattern. The shifted spectral components are recomposed by the second lens and grating into a sigle optical beam and input into the channel. The receiver end consists of a very similar scheme as that of the transmitter, but it uses a conjugate phase mask that reverses the phase shifts of the transmitter.

When the codeword is the phase conjugate of the phase mask used by the receiver, the correlation is maximised (since that means that the receiver is using the same codeword as the transmitter) and the original ultrashort coherent pulse is reconstructed. If another codeword that does not correspond with that of the receiver's enters the decoder, then a pseudorandom burst noise is produced. In this way, other users' tranmissions are ignored.

The presented system implies certain technical difficulties [48]. Among these difficulties we have the obtention of coherent ultrashort light pulses, the base of this model, the correct spatial alignment of all the bulk-optical components, the vulnerabity of the encoded signals to disperson and non-linearities, etc. Certain improvements in miniaturisation by means of InP-based integrated AWGs have been proposed in [63, 64]. A practical implementation

of [62] was presented in [65, 66]

The other main coherent OCDMA technique is Temporal Phase Encoding (TPE), which consists in manipulating the phase of the optical signal in the time domain. TPE OCDMA was first presented in [67, 68]. Temporal Phase Encoding is achieved by producing an ultra-short coherent pulse by a mode-locked laser. This pulse is split into n pulses which suffer different delays and those delayed pulses are then phase-shifted in the time domain according to the user's codeword. The phase shifts in the time domain for the encoder are denoted as $\{\phi_{E1} \cdots \phi_{En}\}$. Finally, the n pulses are combined and a string n consecutive pulses is injected to the network. In the receiver side, a complementary set of phase shifts in a time reversed order $\{\phi_{D1} \cdots \phi_{Dn}\}$ is appplied to the received codeword prior to a delay and recombination of the signal to calculate the autocorrelation function. If the codeword matches, a high autocorrelation peak is produced if not, only noise can be seen at the output of the decoder.

A experimental setup for a tunable temporal phase encoder and decoder usign optical waveguide integration was presented in [67].

OCDMA Network architectures

Code Division Multiple Access (CDMA) [48, 69] techniques have proven to offer higher capacity than Wavelength-Routed Networks (WRN) thanks to the statistical multiplexing properties they offer [70]. Indeed, the use of orthogonal codes allows multiple users to simultaneously transmit on the same frequency band without interfering with each other [48, 69]. However, OCDMA suffers from Multiple Access Interference (MAI), which often arises when more than a given number of users access the shared media simultaneously. This may happen both when we have pseudo-orthogonal (incoherent OCDMA) and theoretically orthogonal (either coherent or incoherent with differential detection) codes. When this occurs, the signal quality drops at the receivers, and all colliding symbols cannot be successfully decoded. This is often referred to as an "outage", and retransmission of all the packets involved in the outage is required. Nevertheless, orthogonal and pseudo-orthogonal codes may achieve high-performance results as long as such MAI limit is not exceeded.

OCDMA techniques have been proposed for access networks (FTTx) [71,72], metropolitan area networks [73–78] and backbone networks [79–82]. For instance, OCDMA codes are very suitable for simplifying the Medium Access Control of the upstream channel in Passive-Optical Networks (PONs). At present, the Optical Line Terminal (OLT) coordinates the access to the upstream channel by the Optical Network Units (ONUs), granting access to them at specific non-overlapping timeslots. OCDMA techniques would allow the users to simultaneously access the channel without interfering each other as long as an acceptable MAI probability limit is not exceeded. In the case of backbone networks, orthogonal codes have been proposed to be used as labels in GMPLS networks [79, 80] on attempts to increase the capacity provided by optical fibres due to statistical multiplexing.

1.3.4 Evolution of Optical Technologies based in Plastic Optical Fibres

Main lines of research in Plastic Optical Fibres

As it was pointed out in Section 1.2.2, Plastic Optical Fibres (POF) have had much less attention than silica core fibres due to their lower bandwidth distance product. On the other hand, POF are less fragile than silica core fibres and are significantly cheaper. Unfortunately, their high attenuation made them pale in comparison with silica core fibres. In the last 20 years, though, there have been significant changes that have improved by leaps and bounds the performance of Plastic Optical Fibres.

The first POF were discovered in 1968 by Dupont [83]. The first POF were based on polymethyl methacrylate (PMMA) and had losses of up to 1000 dB/Km. Dupont's patents were sold ten years later to Mitsubishi Rayon in 1978. Mitsubishi Rayon managed to reduce those loses to 150 dB/km in the 650 nm window by using step index fibres.

Nevertheless, it was professor Yasuhiro Koike who provided the next major step in POF research [84, 85] in 1990 with the Graded Index PMMA (GI-PMMA) POFs. In this contributions they got a 3 GHz-km bandwidth distance product with a loss of 150 dB/km. In 1995, Koike provided another major step by presenting a graded index fibre which was based in perfluorinated polimer. This new POF had a loss of 50 dB/km in the 650-1300 nm band. Another great step in POF research was taken in 2001, when microstructured polymer optical fibres were created.

In 2005, Fuji Photo Film commercialised a 30m POF link at 1 Gbps using a 780 nm Vertical-Cavity Surface Emitting Laser (VCSEL) over a PMMA GI-POF. That same year PMMA GI-POF were introduced to the market by the Optimedia Company of Korea. As a result of all the previous advancements, there have several research projects funded by the European Union in the field of POFs, namely POF-ALL and POF-PLUS. A summary of the goals achieved in the POF-ALL project can be seen in [86].

The POF-ALL project ended in 2009 and addressed research in POF transmission for short and medium range systems. In the medium range, the POF-ALL project managed to areach 100 Mbps over 200 m of standard Step Index (SI)-PMMA-POF, suitable for environments in which a Fast Ethernet connection could be needed. In the short range, the POF-ALL project achieved 1 Gbps over 50 m of GI-PMMA-POF. This could be useful for very short links inside a house to interconnect devices with high bandwidth requirements in short distances (eg: a TV, to receive very high quality video signals).

The POF-PLUS project ended in 2011. Their last major advancements in POF were presented in [87]. This project achieved several goals in optical POF transmission. The first one was 5.8 Gbps over 50 metres of Single-Core GI-POF employing 2-PAM and the help of an algorithm called Decision Feedback Equalisation (DFE). The POF-PLUS project achieved 5.3 Gbps over 50 metres of GI-POF. Furthermore, the POF-PLUS project managed research in the field of multi-core POF fibres. In this sense, the project also produced a 4.7 Gbps transmission over 50 metres of SI multi-core POF. Finally, 10 Gbps over 25 metres of SI multi-core POF were achieved.

Due to the great development of POF in the last decades, POF has undergone a process of standardisation, which has resulted in the VDE 0885-763-1 Standard for High Speed Communication over Plastic Optical Fibers, published in 2012. This standard provides a slotted link-layer definition. This standard is briefly reviewed in the following section.

A brief review of VDE 0885-763-1 standard and its energy efficiency mechanisms

The physical layer specified in the VDE 0885-763-1 standard is based on the parameters shown in tables 1.2 and 1.3 and on the periodic frame structure shown in Fig. 1.19. The parameters in table 1.2 correspond to a sampling frequency (F_s) of 312.5 Megasamples per second (Msps) while those in table 1.3 correspond to a sampling frequency of 62.5 Msps. The contribution of this thesis focuses on the 1000 Mbps physical bit rate case.

The columns of tables 1.2 and 1.3 show, from left to right, the spectral efficiency, the M-ary Pulse Amplitude Modulation used, the number of bits per modulation dimension (three in total), the physical bit rate and the Physical Coding Sublayer. MII stands for Media Independent Interface. GMII and XGMII stand for Gigabit and Ten Gigabit MII respectively. Although this might seem confusing to the reader, the MII indicates bit rates lower than (approx) 100 Mbps, the GMII indicates bit rates between (approx.) 100 Mbps and below 1 Gbps and the XGMII indicates rates between 1 Gbps and 10 Gbps. This contrasts with other technologies like Ethernet, in which there is generally only binary speed for the 100 Mbps, 1 Gbps and 10 Gbps cases.

η (bits/s/Hz/D)	M-PAM	nb(1) (bits/dim)	nb(2) (bits/dim)	nb(3) (bits/dim)	PHY bit rate (Mb/s)	PCS interface
0.8254	2	1	0	0	249	GMII
1.3145	4	1	0.5	0	396	GMII
1.8145	4	1	1	0	547	GMII
2.3145	8	1	1	0.5	698	GMII
2.8145	8	1	1	1.0	849	GMII
3.3145	16	1	1	1.5	1000	GMII
3.8145	16	1	1	2.0	1150	XGMII
4.3145	32	1	1	2.5	1301	XGMII
4.8145	32	1	1	3.0	1452	XGMII
5.3145	64	1	1	3.5	1603	XGMII
5.8145	64	1	1	4.0	1754	XGMII

Table 1.2: VDE 0885-763-1 modulation parameters for $F_s = 312.5 Msps$

Regarding the frame structure, it contains pilots, physical layer headers and codewords [88]. Pilots (S1,S2) are used for physical layer functions like timing recovery and equalization. Physical Layer Headers (PHS) are used to exchange parameters between the edges, for instance coefficient equalizers. Finally, the codewords (CW) are used to carry the user data bits along with extra bits for error correction. The VDE 0885-763-1 standard uses Multilevel Coset Coding (MLCC) [89] with three coding levels. Different Bose-Chaudhuri-Hocquenghem (BCH) codes are used for error correction in each level.

As shown in Fig. 1.19, each physical layer frame comprises one S1 and 13 S2 pilot sub-blocks, 14 physical headers sub-blocks and 112 codewords (CW). Each codeword comprises 2016 symbols, while sub-block pilots and headers contain only 160 symbols.

Thus, each physical layer has a total of:

$$112 \times 2016 + (1 + 13 + 14) * 160 = 230272 \text{ symbols}$$

η (bits/s/Hz/D)	M-PAM	nb(1) (bits/dim)	nb(2) (bits/dim)	nb(3) (bits/dim)	PHY bit rate (Mb/s)	PCS interface
0.8254	2	1	0	0	49	MII
1.3145	4	1	0.5	0	79	MII
1.8145	4	1	1	0	109	MII
2.3145	8	1	1	0.5	139	GMII
2.8145	8	1	1	1.0	169	GMII
3.3145	16	1	1	1.5	200	GMII
3.8145	16	1	1	2.0	230	GMII
4.3145	32	1	1	2.5	260	GMII
4.8145	32	1	1	3.0	290	GMII
5.3145	64	1	1	3.5	320	GMII
5.8145	64	1	1	4.0	350	GMII

Table 1.3: VDE 0885-763-1 modulation parameters for $F_s = 62.5Msps$

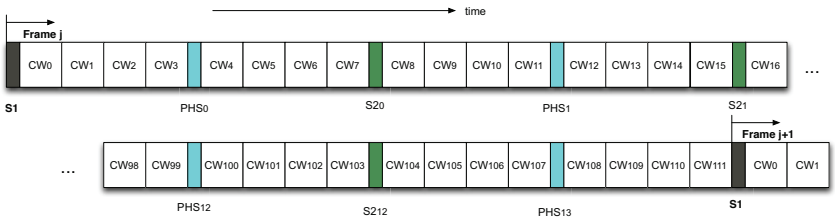


Figure 1.19: Illustration of the frame structure in the VDE 0885-763-1 standard

When using the 1 Gb/s configuration, the symbol frequency is 312.5 Mhz, thus the transmission of a frame requires $736.8704\mu s$.

Layer-2 data frames are transmitted back-to-back over groups of 4 CW. Each group of 4 CW + PHS/S then takes:

$$\frac{(4 \times 2016 + 160 \text{ symbols})}{312.5 \text{ MHz}} = 26.3168\mu s \text{ per group}$$

At 1 Gbit/s configuration, such a group can carry 26316.8 layer-2 bits (approximately 3290 bytes) as it follows from:

$$\frac{26.3168\mu s}{10^9 \text{ bit/s}} = 26316.8 \text{ bit}$$

Another important point of the the VDE 0885-763-1 standard is that it enables energy savings by stopping the transmission of groups of four codewords as illustrated in Fig. 1.20. In the absence of data, the link enters the low-power mode and only the pilots and physical headers are transmitted to ensure that the transmitter and receiver are fully aligned and ready to begin data transmission as soon as new data arrives for transmission.

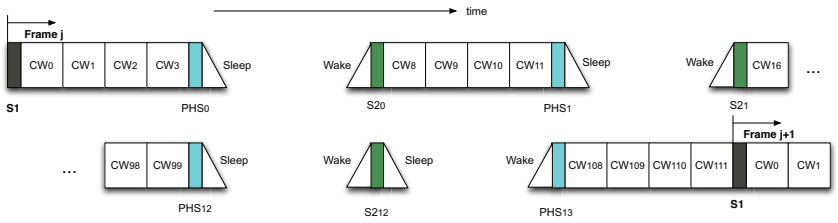


Figure 1.20: Illustration of the use of the low power mode defined in the VDE 0885-763-1 standard

If a new packet arrives during the low-power mode, the standard states that the transceiver must wait until the next group of codewords to activate the link and dispatch such packet.

This is rather different than the IEEE 802.3az Energy Efficient Ethernet (EEE) standard, whereby a given link may be activated or deactivated at any time. Furthermore, in EEE, the transition times from sleep to active (wake up times T_w) and from active to sleep (sleep times T_s) are substantially large, thus producing an important energy penalty if many transitions between states occur (see [4]). This effect is particularly harmful for short packets and at low traffic loads.

1.3.5 Storage Technologies

SAN vs NAS

Storage is an essential resource in today's Information Society. The ever-increasing need for storage has resulted in the development of new technologies to obtain higher density hard disks and optical disks. All the more important as our so called information society continues to develop. While it seems that for the average end user higher storage density media is a

reasonable solution, it is hardly the case for medium to large enterprises. This problem has led to the development of non-local storage technologies and solutions to separate the computational elements (eg: a PC or a server) of an IT (Information Technologies) system from its storage component (eg: a hard disk).

There are two main paradigms to achieve this separation. These two paradigms are complementary rather than mutually exclusive. That is, one can have a mixed scenario in which both paradigms are implemented and work in conjunction. To introduce both, let us consider Fig. 1.21.

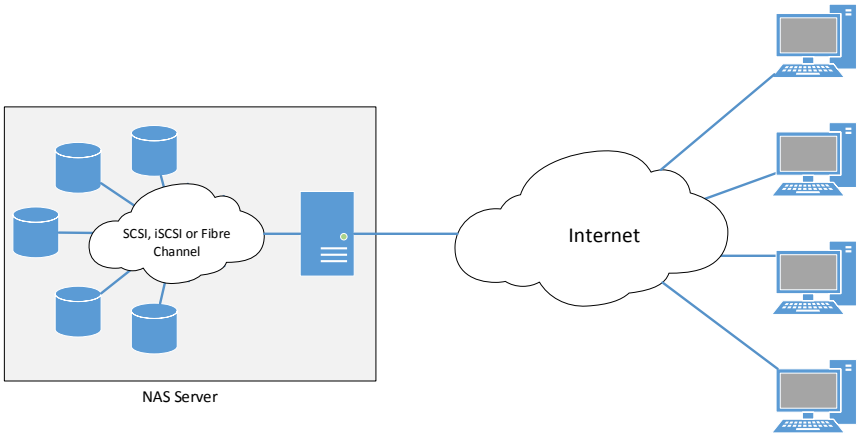


Figure 1.21: Figure of a NAS and SAN environment, inspired by Fig. 4.6 in [1]

One of those paradigms is what is known as Network Attached Storage. Network Attached Storage is a file server specifically designed and prepared to serve files through a network. Network Attached Storage is made possible by technologies that allow file-level access through a data network [1]. Among the main technologies that represent this paradigm we have remote filesystem technologies, like NFS and SMB/CIFS. Network Attached Storage allows file access through a network using the file as the basic unit for such access. Access through remote filesystem technologies is, in the interface aspect, transparent to end users, being only in performance (delay, bandwidth) that there is an apparent difference from a usage point of view.

The other paradigm that is represented in Fig. 1.21 is what is known as Storage Area Networks. One of the most clear differentiation points between Storage Area Networks and Network Attached Storage is that Storage Area Networks work with storage systems in a block-oriented mode rather than in a file oriented mode. That is, instead of writing files, they deal directly with individual blocks of the final storage medium. This paradigm is closer to the physical medium than NAS, where a NAS client does not know (and does not need to know) which is the server's filesystem for the final physical medium. In Fig. 1.21, SANs are represented in a generic cloud with several discs attached. The technology in the cloud can be simple SCSI (if the storage medium is attached directly to the server), iSCSI or Fibre Channel (in the last two, if the storage medium is not attached to the server). Although the SAN part does not belong to the NAS server itself, it is represented inside because the SAN part is presented to the Operating System as another block device just like a normal SCSI device would be.

There are other features of SANs that are commong among some of the technologies integrating this paradigm but are not defining but not common to all of them. One of them is that some SAN technologies have their own specialised (and even exclusive) network infrastructure and protocol stack. Others simply reuse existing networks and protocol stacks or take intermediate solutions between both extremes.

NAS technologies

NAS servers are specifically engineered computers to the sole purpose of providing file access through a network. They usually sport fixed configurations and in some occasions minimal and/or specialised operating systems to improve file serving performance. To better serve their purpose of sharing files quickly they are usually connected to a high speed LAN (eg: Gigabit Ethernet). Another common feature of NAS solutions is that they are usually designed to be Plug and Play, so the initial configuration and later maintenance are minimal [1].

In providing NAS services, network filesystems are a key component of NAS servers. Among the most representative remote filesystems we have the Network File System (NFS) and the Server Message Block/Common Internet File System (SMB/CIFS).

Network File System (NFS) is a distributed file system protocol originally developed by Sun Microsystems [90]. It is currently standardised by the IETF and the current specification is NFS v4.1 defined in RFC 5661 [91]. It implements the Remote Procedure Call Standard defined in RFC 5403 [92]. NFS v4.1 provides several improvements over prior versions [91], being the main two important performance improvements and a stateful protocol definition. NFS is a filesystem usually supported in Unix (Solaris, MacOS X, HP-UX, etc) and Unix-like (Linux, FreeBSD, etc) operating systems.

SMB/CIFS [1,93] is Microsoft's version of a remote filesystem it is mainly implemented on Windows, but also on some Unix systems [93]. It is usually implemented over NetBIOS over TCP/IP. It provides functionality for session control and file access. Like NFS, it provides a transparent interface for users, which see the SMB/CIFS filesystems as any other filesystem present in the machine. Due to being a proprietary standard, its specifications are far less accessible than those of NFS.

However, NAS systems are not the definitive remote storage solution. NAS servers usually implement application protocols to emulate a filesystem on top of (usually) the TCP/IP stack. This implies several bottlenecks because of how remote filesystems work and because of the protocol stacks required for the implementation [1].

Regarding remote filesystems, there is a clear bottleneck in the server due to the necessity of taking the data from the local hard disk, copying it through the PCI bus to the Operating System's memory, processing it by the CPU. It also includes the processing related to the application protocol implemented (NFS or SMB/CIFS). After all this has happened at the OS level, then the data is encapsulated in the TCP/IP or equivalent protocol stack and delivered to the network card through the PCI bus again. Apart from the data traversing the PCI bus twice, it can be seen that this paradigm is very CPU intensive due to relying on the OS regulating the access to the local filesystem for the remote filesystem clients, as well as the overhead brought to the table by the TCP/IP stack itself.

Because of these reasons, NAS technologies are not adequate for intensive I/O systems,

like large databases [1].

Brief introduction to SAN technologies

A Storage Area Network (SAN) is a dedicated network that provides access at a block level to a remote storage [1], mainly mapping SCSI over a network protocol stack. This is a major difference with Network Attached Storage, which works at a file level rather than block level. SAN technologies send direct read and write block operations over the network. Furthermore, while NAS are usually seen as file servers by the operating system, SANs are represented to the operating system, usually by means of special drivers, as SCSI disks.

Most families of technologies related to SANs are a derivative of Fibre Channel (FC) [1]. A Fibre Channel has its own dedicated network infrastructure, protocol stack and physical medium. While this is most convenient for performance issues, it also comes at an important economic cost. It is for this reason that other members of the Fibre Channel family (and other SAN technologies in general) try to reduce those costs by doing a hybrid of Fibre Channel and other existing technologies.

Another dedicated technology used for SANs is InfiniBand, which replaces the PCI bus with a serial network integrated by InfiniBand switches [1]. Although it is presented here as a SAN technology, it would be more appropriate to label it as a high-speed network technology to connect computing resources, which means it is used for high performance computing, making it a technology of a wider scope than simply provinding remote access to storage. Due to its more generalistic nature, InfiniBand will not be covered with much detail in this thesis.

On the opposite side of SAN technologies, there is iSCSI. iSCSI does not require a dedicated network infrastructure, since it is simply a protocol to transport SCSI over TCP [94]. This implies that it can be implemented at a software level in any computing device.

iSCSI

The Internet Small Computer Systems Interface (iSCSI) is a technology initially defined by Cisco and IBM and initially standardised in 2004 through the IETF [95] and later updated in 2014 [94]. iSCSI fully adapts the SCSI architecture onto the TCP protocol stack. Because of this, it can be implemented on any computing device, leading to the appearance of both proprietary and free [96] implementations.

iSCSI defines a server-client architecture in which a SCSI client is denominated an “Initiator”. Initiators send SCSI commands to request access to SCSI resources managed by a server. Those SCSI resources are called “targets”. The iSCSI initiators receive responses from the iSCSI servers according to the request-reply nature of the SCSI protocol. SCSI devices are called “Logical Units” [1], which are identified by Logical Unit Numbers (LUNs).

While it can be implemented in many computing devices, the fact that iSCSI rests on the TCP/IP stack is also its highest drawback. As in it happens with NAS technologies, the TCP/IP protocol stack is not the best suited for SAN technologies due to its heavy nature. Other technologies, like the Fibre Channel family technologies, provide better suited protocol stacks for SAN traffic.

Fibre Channel technologies

Fibre Channel (FC) [1] is a high-speed network technology, which runs commonly at 2, 4, 8 and 16 Gigabits per second. It defines a full protocol stack to connect computers to remote data storage. It defines, following the OSI model, from the lowest parts of the stack (physical, cabling and modulation) to the highest (protocol mapping layer). The original FC protocol stack can be seen in Fig. 1.22. Other architectures over which the Fibre Channel Protocol can be stacked are also represented in Fig. 1.22.

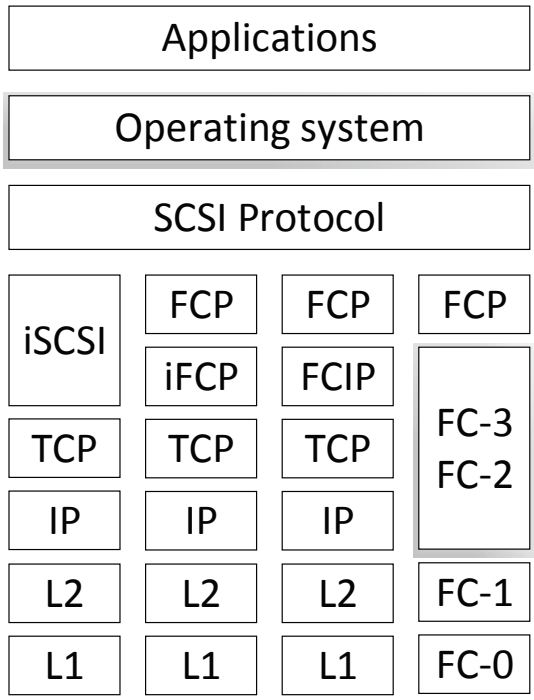


Figure 1.22: Protocol stack for the different technologies of the Fibre Channel family, inspired by Fig. 3.39 in [1]

The lowest level of the FC protocol stack is FC-0, which defines cables, plugs and signal encoding [1]. By contrast with SCSI buses (which are parallel), data over an FC network is transmitted serially. This requires that FC bitrates compensate for the lack of parallelism with higher bitrates. FC components can be Base2 or Base10. Base2 components have a bitrate which is a multiple of a power of 2 Gbps (including the case of 1 Gbps). Base10 components have bitrates which are multiples of 10 Gbps and have to be compatible with one prior generation, except for the 100 Gbps standard [1]. The FC standard demands that the bit error rate is 10^{-12} or lower, and is usually defined for either multimode or single mode fibre, but the standard. was also extended to support copper (with, of course, a shorter range). [1]

The next level of the Fibre Channel Protocol stack is FC-1, which mainly describes the line coding (8b/10b in most ocassions and 64b/66b in specific cases), the definition of ordered sets (transmission words) and the link control protocol. The FC-2 level defines how data transfer takes place in the FC network: it regulates flow control and defines service

classes for the different applications. Flow control is double in FC. The first flow control is the end-to-end flow control, which is among end devices and is based on a negotiated credit system. The second flow control is defined on a per-link basis. In this per-link flow control, a buffer-to-buffer credit between ports is negotiated [1]. Regarding traffic classes, there are three main classes for data traffic. Class 1 defines a connection oriented communication which reserves all the bandwidth for it and guarantees ordered delivery. Classes 2 and 3 are packet-oriented, so several of them can exist simultaneously. Class 2 provides end-to-end and per-link flow control as well as ordered delivery (if needed), while Class 3 only provides only per-link flow control, leaving frame loss detection to higher layers.

FC-3 layer provides common services and is said to possibly provide functions for multipath, data encryption and compression among others. Unfortunately, though, this layer still remains to be fully defined. To the moment of writing this thesis, FC-3 is being defined to provide what is called Basic Link Services [97] and Extended Link Services [98]. Nevertheless, it has been for long a layer almost empty of content whose services were provided at some other layers [1].

FC-4 is the layer that maps the application protocols to the FC SAN. FC-4 is implemented just below the SCSI API and is presented to the higher levels as an alternative communication protocol known as Fibre Channel Protocol (FCP) that provides an equivalent functionality to that of other SCSI protocols, such as the Parallel SCSI. FCP is implemented as an additional driver at the operating system and devices accessible through it are mapped as common SCSI devices. This way, an almost transparent mapping is provided at the client level [1].

Nevertheless, the main issue with traditional FC is that it requires a specialised network of its own. This leads to having a separate network infrastructure just for Fibre Channel. Apart from being not very cost effective, there is also the need of interconnection of storage networks over long distances. Regarding the interconnection of SANs over long distances, the Fibre Channel over IP protocol (FCIP) was defined as a means of connecting FC SANs through an IP cloud [99]. FCIP interfaces are defined at the edge of FC SAN islands, where FCP frames are mapped onto TCP/IP connections and sent through an IP network. At the other side of the network, the corresponding FCIP interface decapsulates the FC frames and introduces them into the second FC island.

By contrast, the Internet Fibre Channel Protocol (iFCP) [100] provides a means of building a SAN on top an IP network. That is, instead of connecting pure FC islands like FCIP does, iFCP intends to allow FC devices to be connected to an IP SAN. As in FCIP, we may of course connect several IP SANs through an IP backbone. This is achieved by providing mixed fibre channel fabrics in which special ports known as FL-ports permit the FC devices to connect to an IP network.

Fibre Channel over Ethernet

Finally, there is Fibre Channel over Ethernet (FCoE), which tries to unify SANs and LANs by encapsulating FC frames over Ethernet frames. FCoE is defined in the INCITS FC-BB-5 (Fibre Channel Back Bone 5) specification [101]. While other approaches try to migrate Fibre Channel to other transmission technologies, FCoE tries to achieve the opposite, that is, to adapt an existing transmission technology to the needs of the higher levels

of FC [1]. To this end, a reliable transmission level is required, something that Ethernet does not provide in principle. FCoE advocates considered why iSCSI, despite its relying on TCP, was not the best choice for SAN traffic. Their main concern is that TCP is a stateful protocol with certain overhead, thus difficult to implement at the hardware level and that there could be other possibilities which could run on simpler protocols more easily implementableon hardware [1]. Furthermore, iSCSI is not compatible with Fibre Channel, which is the solution with higher acceptance, something that can be clearly seen in the wider range of technologies that try to support FC, many of them mentioned above.

Companies have traditionally run two separate networks in their datacentre environments: An Ethernet-based LAN for data communications and Internet access with typically 100/1000-Mbps Ethernet Network Interface Cards (NICs), and a Fibre Channel storage area network (SAN) with 1/2/4-Gbps Host Bus Adapters (HBAs) for storage connectivity [102]. Obviously, such a duplicated network architecture requires more equipment and higher Operation, Administration and Maintenance (OAM). Consolidation of the two networks into a single Ethernet-based one, is expected to bring large cost reductions while reducing the OAM.

To this end, the IEEE has created the Data Centre Bridging (DCB) task group to adapt Ethernet to the high-performance lossless features required in datacentre scenarios. The DCB task group involves several IEEE standards: Priority-based Flow Control (IEEE 802.1Qbb) [103], Enhanced Transmission Selection and Data Centre Bridging Exchange (defined in IEEE 802.1Qaz) [104], and Congestion Notification (IEEE 802.1Qau) [105]. A good summary of their role in the definition of such an Enhanced Ethernet can be found in [106].

Normally, packet loss in Fibre Channel fabrics is avoided by implementing a point-to-point, credit-based flow control system. In this kind of control, the downstream node of the link controls the credit given to the upstream node by decreasing a credit counter whenever it receives traffic and increasing it when it transmits the traffic it receives. In a similar fashion, the upstream part of the link takes into account the traffic it may send and decreases its credit every time it sends a packet. The credit is increased by the upstream node every time it receives a packet. If the upstream node runs out of credit, it stops sending packet until it gets the credit necessary to transmit it.

On the other hand, Ethernet has traditionally been a lossy network, where congestion was treated by discarding packets when buffers overflow. Flow and congestion control was left to higher layers to reduce the cost and complexity of Ethernet fabrics. Most of this functionality relied on the TCP/IP protocol stack, which has been the most used stack for network protocols. In the TCP/IP stack, packet loss is taken as a sign of congestion, leading to retransmission and a reduction in the sender's transmission rate. Nonetheless, this behaviour leads to lower than desired performance in modern datacentre scenarios and this has led to Ethernet, among other technologies, to try to bring the flow and congestion control back to layer 2 of the OSI stack. It is worth noting that there had been previous initiatives to bring lossless operation to Ethernet, but it had been rarely used in the past due to the predominance of the TCP/IP stack. With the new need of joining LANs and SANs, the Data Centre Bridging (DCB) group at the IEEE aims at providing lossy and lossless operation in the same network at the same time for different services.

In this sense, the IEEE 802.1Qbb standard (approved in 2011) defines a per-link Priority

Flow Control mechanism based on the basic IEEE 802.3x PAUSE semantics. Such a PAUSE frame allows the receiving switch to communicate buffer availability to the sender switch for each traffic class on attempts to avoid packet loss due to buffer overflow causes. Essentially, when the receiver's buffer achieves a certain buffer threshold, it transmits a PAUSE frame to the sender, which stops transmitting data packets. When the buffer empties to a certain level, then an UNPAUSE message is sent, so the transmission starts again. In a nutshell, it is a protocol aimed at regulating link activity. By contrast with traditional FC per-link flow control, PFC is based on an XON/XOFF paradigm, rather than a credit system. The main difference between Priority Flow Control and IEEE 802.3x PAUSE is that it distinguishes traffic of different priorities.

It can be seen here that two thresholds are defined and that their design highly depends on the Round Trip Time (RTT) of the link. Those thresholds must be designed so there is no underflow (in the case of the lower threshold) or overflow (in the case of the upper threshold). XON and XOFF signals must be sent according to them to maximise link utilisation. One of the main drawbacks of XON/XOFF control is that its reaction time is 2 RTT, so buffers must be sized accordingly, by contrast with a credit-based system. The XON/XOFF system requires 1 RTT to send the message from the downstream node to the upstream node and another RTT until the effects of this message are noticed at the downstream node since, even if the processing time is negligible, no packets will arrive (or cease to arrive) until another RTT since it requires an RTT for the first packet to arrive (or for the last packet to arrive). By contrast, a credit-driven control system only requires buffer dimensioning for RTT seconds since credit updates take RTT seconds to arrive. Nevertheless, there is one advantage to the XON/XOFF paradigm and it is that it has less signaling overhead, but has another disadvantage and is that, while losing a credit update can lead to lower link utilisation during a certain interval of time, losing an XOFF signal can lead to buffer overflow if such a frame is lost.

These issues were subsequently studied in [107], where a detailed analysis of PFC is made for Cisco Nexus 5000 switches, comparing it with credit-driven flow control. There are five main conclusions to be taken from this study.

- Both the credit system and PFC system are limited by distance.
- While increasing the distance above the maximum supported threshold leads to low link utilisation in both cases, it is also true that in the case of PFC there is also a distance above which also the lossless semantics are lost since, for a given buffer size, both lower and upper thresholds overlap completely.
- There is a maximum distance associated to a certain buffer size.
- The dimensioning of buffers is made over the hugely conservative assumption that buffers fill at the full link rate and, at the same time, they are not emptied at all.
- An indirect conclusion is that, if we were to increase the maximum reachable distance for an FCoE network, it might lead to huge buffer sizes to guarantee lossless operation.

Especially the last conclusion is to be taken into account since there is much room for improvement if we take into account that not always links operate at full speed and that

also links empty at a given rate, which is normally greater than zero. In a nutshell, if we take into account statistical properties of the traffic we might see that, for normal operation, current buffer sizes might guarantee a reasonable frame loss probability even if there is no flow control.

While PFC controls congestion at a link level, this control has only a local scope, but ignores the more global picture of the network. If persistent, congestion in a single link can propagate to the previous links, leading to that congestion being noticed across a whole tree in the network to the very end nodes that originate traffic through that link. The thing is that, not all of these nodes contributed in the same significant way to the congestion of the network. This originates a *head of the line blocking* problem, which means that traffic not directed to the congested part of the network is unfairly blocked because there is congestion several links ahead. In this sense, lossless flow control in a link can be a source for congestion too.

To complement local flow control, the IEEE 802.1Qau Congestion Notification [105] standard defines Congestion Notification to source and end stations. This standard provides the following functionalitie:

- Congestion detection: this is made by monitoring the egress queues of the switches and detect congestion.
- Congestion notification: this is achieved by implementing a protocol that sends back a unicast frame to the station source of the congestion.
- Congestion response: the ingress node responds accordingly when receiving a congestion notification frame.

Each bridge on the way to the final destination define what are known as Congestion Points (CP) at its output queues. The source end station is known as a Reaction Point (RP). CPs send Congestion Notification Messages (CNM) back to the source of a frame is congestion is detected. The source then adapts its output rate according to the messages received following an algorithm known as Quantized Congestion Notification (QCN) Protocol [108]. This algorithm works in a similar fashion as the TCP congestion control.

After receiving a CNM message, a rate limiter is triggered to control the output rate of the ingress node that receives the message. Each time a CNM message is received, a multiplicative decrease of the input rate is performed at the Reaction Points. In the absence of CNM messages, the ingress node tries to recover the original transmission rate. This is made in three stages, known as fast recovery, active increase, and hyperactive increase respectely. Each stage increases the rate faster than the previous one. Having multiple stages allows for a quick search and stabilisation at an appropriate rate for the network status, leading to share of bandwidth at the Congestion Point.

Finally, Enhanced Transmission Selection (ETS) [104] defines multiple priority groups which can be used to separate, among others, LAN and storage, guaranteeing a minimal bandwidth for each priority group.

In the next section, the problems that have arised due to the need of extending the reach of SANs are explained, as well as the solutions taken to address those problems.

The ever increasing need (and problem) of distance

The development of the aforementioned technologies has been fostered by the need of extending the reach of SANs. This reach extension has also become a problem as well as a need due to the traditional short distances of storage communications.

The concept of "distance as a need" arises from the fact that data survivability is becoming a matter of life and death for many companies. In a world with full computerisation, a company can dissappear overnight due to a massive loss of information if their files are fully centralised. Consistency is a need and this need for consistency requires a certain degree of centralisation in many cases. Unfortunately, this also poses an important vulnerability on storage since this leads to a single or few points of failure. Losing information for a bank means not being able to know anymore who are your customers or, even worse, what money do they have or they owe your company. This is clearly a problem posed by the use of computerised storage in favour of physical information.

Much of the work related with the extension of SANs is related to WDM networks-SAN integration, especially in the WDM MAN ring scenario. In [109], an $FT - FR^4$ transparent WDM MAN ring architecture of around 138 km long is suggested and evaluated through simulation under Poisson and self-similar traffic to interconnect SANs. This model was later changed and extended by the authors in [109] in a later contribution [110,111]. In [110,111], a sectioned WDM ring similar in length to that of [109] is suggested as a possible SAN extension scenario for data protection against catastrophes. In [110, 111], both $FT - FR$ and $TT - FR$ are evaluated. Further WDM ring and SAN integration with different slotting algorithm for WDM rings was investigated in [112]. The architecture in [110, 111] was later used to suggest a synchronous mirroring scheme for catastrophe survivability in SANs under optical WDM rings in [113]. Finally, an extension scenario in a mixed Optical Circuit Switching (OCS)/Optical Burst Switching (OBS) mesh network was presented in [114].

There have also been great efforts regarding the study of iSCSI in WAN scenarios. Authors in [115] measured the performance of iSCSI for several network delays and block sizes. Authors in [116] further studied and modeled the performance of iSCSI in MAN and WAN networks on distances of up to 1000 km and several packet loss probabilities. Both works show that iSCSI is clearly affected by loss probability and, especially, by distance. This is natural since iSCSI rests over TCP/IP and TCP throughput is affected by distance due to the congestion avoidance algorithms that regulate its behaviour.

In this sense, several efforts have been posed to improve iSCSI throughput in the WAN. In [117], steps to tune the SendBuffer of Open iSCSI in order to improve throughput are explained. Additionally, some crosslayer solutions, like that of [118] in which an interaction between the TCP and iSCSI layer to improve throughput is suggested. Other solutions suggest to change the common TCP layer for other implementation [119] or even creating a version of TCP especially aimed towards datacentres, such as Data Centre TCP (DCTCP) [120].

Other efforts have gone in the direction of using several parallel connections to improve packet throughput, as is the case of [121], where an architecture for parallel processing of iSCSI packets was suggested. The authors in [122] later proved that certain throughput improvement could be achieved by using up to 4 parallel TCP connection on iSCSI. Finally, the authors in [123–125] presented a system capable of improving iSCSI throughput in WAN environments by automatically tuning the number of multiple connections of the iSCSI session.

1.4 Contributions of this thesis and progress beyond the State of the Art

Once reviewed the state of the art, the reader has a perspective of the general progress of each of the three main areas of technologies covered in this thesis. It is time to outline now the four contributions in this work and link them to the technologies studied in the state of the art:

- Contribution C1 studies the integration of transparent WDM rings with Metro Ethernet in chapter 2. In this chapter, an adaptation box to integrate Ethernet and TT-FR WDM rings is proposed. In addition, two broadcast and multicast strategies are studied that take advantage from the circularity of a WDM ring network. This contribution clearly covers the Optical and Ethernet blocks displayed in Fig. 1.1 through WDM and Ethernet technologies. This contribution has been published in the IEEE Network Magazine.

- Contribution C2 covers the design and analysis of hybrid WDM-OCDMA rings in chapter 3. In this contribution, a mixed WDM-OCDMA architecture is proposed. The ring is divided into several segments in which there is a hub node that takes the traffic from the rest of the nodes in the segment, called regular nodes. Regular nodes forward their traffic transparently to the hub through OCDMA. Hub nodes forward the aggregated traffic from the Regular nodes in their segment to other hub nodes through WDM wavelengths. Then, hub nodes deliver traffic to Regular nodes in a segment through OCDMA. This contribution covers the Optical area displayed in Fig. 1.1 through the WDM and OCDMA technologies. This contribution was presented at the IEEE HPSR 2011 conference.

- Contribution C3 covers the design and analysis of buffers for FCoE in the presence of bursty traffic in chapter 4. In this contribution, a burst arrival Poisson process is used to analyse the buffer overflow probability we would obtain if there were no Priority Flow Control. Our results are further reinforced by simulation with synthetic traffic and generated FCoE traces. This contribution shows that current buffer dimensioning is overly conservative and that, if statistical properties of the traffic involved are used, futher improvements can be made in designing buffers for FCoE switches. This contribution convers the Ethernet and Storage areas mostly through the study of Fibre Channel, FCoE and Convergence Enhanced Ethernet (CEE). The results of this contribution were published in IEEE Communications Letters.

- Contribution C4, in chapter 5, goes mostly back to the Optical Technologies block and draws inspiration from advances in Energy Efficient Ethernet, which belongs to the Ethernet block. This contribution reviews the latest advancements in Plastic Optical Fibres and the recent standard for Plastic Optical Fibes known as the VDE 0885-763-1. In this contribution, two coalescing strategies are proposed to improve energy efficiency in Plastic Optical Fibres transmission following the mentioned standard. Our conclusions are supported by simulation with synthetic traffic and Ethernet traces obtained from different environments. This contribution shows that significant energy

savings are achievable thanks to coalescing and that a reasonable trade off between savings and delay is possible. Our results have been submitted to the IEEE Journal of Optical Communications and Networking.

The reader may realise the heterogeneity of this thesis's contributions. Mainly, a PhD thesis might focus in specific ways of analysing the challenges presented by the thesis, that is, using a specific methodology to solve the problems to be solved. A remarkable fact about our research has been that we have studied several network-related technologies adopting the most adequate tools in each article depending on the nature of the challenge to be solved. The set of necessary skills in each contribution can be seen in Fig. 1.23:

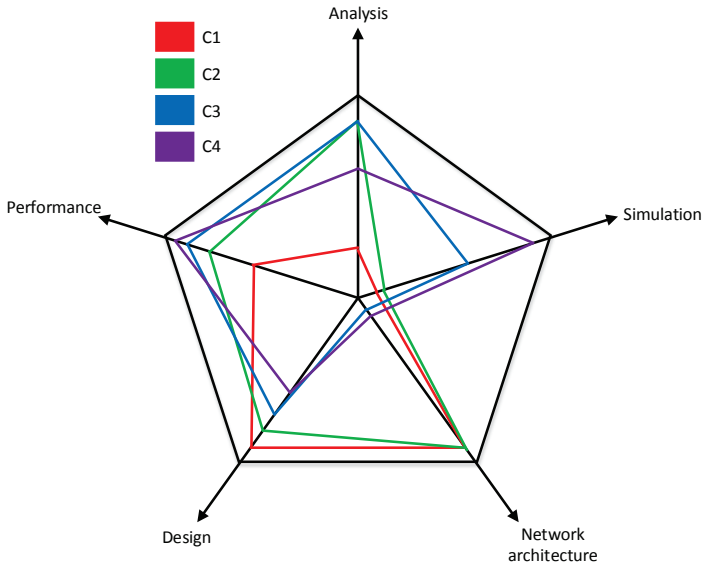


Figure 1.23: Classification regarding the skills involved in each contribution of the thesis.

- Analysis: One way of studying network technologies is by analysing traffic through mathematical models. This has been the main path followed in C2 and C3. There is some analysis, to a certain extent, also in C1 and C4, but not at the same level of depth and complexity, being C1 the less analysis-oriented. Mainly, C2 and C3 model network traffic by means of queueing theory. One of the most important performance metrics evaluated is buffer overflow probability in C3.

- Design: there are always performance metrics that constrain our design. In contribution C1, network design to adapt Metro Ethernet to WDM rings is explored. In contribution C2, OCDMA/WDM ring network design is explored through the constrain of BER given by the Multiple Access Interference of OCDMA. In contribution C3, buffer design in the presence of bursty traffic of FCoE environments is investigated. In contribution C4, we get into algorithm design to improve energy efficiency in Plastic Optical Fibres following the frame format imposed by the VDE 0885-763-1 standard. It can be seen that C1 focuses heavily on design, although contributions C2, C3 and C4 also take this aspect into account.

- Network architecture: In contributions C1 and C2, different network architectures are explored. In C2, as previously stated, a hybrid OCDMA/WDM ring is considered. C1 considers the integration of transparent WDM rings and Metro Ethernet. Contributions C3 and C4 do not enter into the Network architecture domain at all, but rather focus on certain metrics.

- Performance: In all this thesis's contributions, performance metrics (delay, buffer overflow probability, energy consumption, etc) are evaluated either through analysis, simulation or both. In C1, we slightly evaluate the bandwidth spent by different multicast and broadcast strategies. In C2, the Multiple Access Interference of a hybrid OCDMA-WDM ring architecture is evaluated. In C3, performance is evaluated through block probability and buffer size to achieve a certain block probability. Finally, in C4 two coalescing strategies are evaluated in terms of energy efficiency for Plastic Optical Fibres. Contributions C2, C3 and C4 are especially oriented towards performance evaluation, while C1 only performs a rough evaluation of the proposed architecture.

- Simulation: Contributions C3 and C4 required the development of simulators to validate either our analysis (C3) or our algorithm design (C4). Simulators allow us to complement our analysis or even get results that might be difficult to achieve through traditional analysis due to its complexity. This means of evaluation was not used in C1 and C2.

1.5 Conclusions and thesis structure

This chapter has reviewed the motivation and scope of this thesis. This motivation has shown the importance of research in optical Metropolitan and Wide Area Networks. This study involves the study of different heterogeneous techniques that can be grouped into three main areas, namely optical technologies, Ethernet technologies and storage technologies. Next, the historical evolution of these three groups of technologies has been briefly reviewed and the latest research on each area has been summarised in a state of the art. Finally, having exposed the challenges identified in each area the main contributions beyond the state of the art have been introduced. Given the heterogeneity of the contributions of this thesis both in terms of technologies and methodologies, we have associated each contribution with the technologies involved and with the skills or tools (analysis, design, network architecture, performance and simulation) needed to correctly address the challenges presented.

The next chapters deeply analyse each contribution in detail. In chapter 2, the integration of Ethernet and TT-FR WDM ring architectures is studied. In chapter 3, a hybrid WDM-OCDMA ring architectures is proposed. Then, chapter 4 discusses the design of a buffer for bursty traffic conditions for FCoE. Next, in chapter 5, two algorithms for packet coalescing in Plastic Optical Fibres are studied. Finally, chapter 6 concludes this thesis with its main contributions, findings, and future work.

Chapter 2

Converged Metro Ethernet and transparent WDM Ring Architecture

2.1 Motivation

The long list of benefits (especially cost and capacity) of Ethernet LANs has made the IEEE, the ITU-T, and the Metro Ethernet Forum define the requirements for taking Ethernet beyond the local area, toward the metropolitan region. In addition, the ever-increasing traffic demands of new applications and users can only be met by the huge bandwidth capacity provided by optical fibers. This chapter studies how to provide metro Ethernet services over transparent tunable-transmitter fixed-receiver WDM optical ring networks. A new adaptation layer of ME to WDM is proposed, and its benefits and drawbacks are studied. It is shown that such a transparent WDM ring network can be seen as a logical full-mesh topology by the upper ME layer, thus reducing to one the number of optical-electronic-optical conversions per unicast frame. Additionally, two different approaches are proposed in the case of broadcast/multicast traffic, since this may bring scalability difficulties in ring topologies.

It is estimated that about 90 percent of total traffic traversing the Internet has been both generated on and destined to Ethernet LANs. Indeed, the success of Ethernet in the local area is mainly justified by its low cost and great capacity features. For these reasons, a number of standardization bodies such as the IEEE, the International Telecommunication Union-Telecommunication Standardization Sector (ITU-T), and the Metro Ethernet Forum (MEF) have put significant effort into devising the requirements to take Ethernet beyond the local area and extend it to the metropolitan region. Such efforts have given rise to the IEEE 802.1ad provider bridges (PB), IEEE 802.1ah provider backbone bridges (PBB), and IEEE 802.1Qay provider backbone bridges with traffic engineering (PBBTE), the latter completed in 2009.

Besides, the research community has also understood that only optical networks can provide the capacity required to overcome the ever-increasing bandwidth demands of users, especially due to the rise of video streaming applications such as IP television (IPTV) and video on demand (VOD). Indeed, optical fibers provide an amount of bandwidth on the order of terahertz, which, thanks to wavelength-division multiplexing (WDM), can provide a potential capacity of terabits per second. In addition, the local loop is beginning to migrate to optical- based solutions by means of fiber to the x (FTTx, either home, cabinet, or build-

ing), to satisfy the bandwidth demands of the next generation of applications. In light of this, the Ethernet passive optical network (EPON) has been proposed as the Ethernet-based optical access solution to aggregate the traffic generated by thousands of end-subscribers up to the central office, which provides the bridge toward the upper metropolitan area network (MAN). At present, most MAN solutions are based on complex and costly time-division multiplexing (TDM)-based synchronous optical network/digital hierarchy (SONET/SDH) metropolitan networks, which typically provide OC192 circuits operating at 10 Gb/sec over an underlying optical ring infrastructure. SONET/SDH networks are limited by the electronic processing bottleneck, and newer all-optical transparent WDM-based networks have already been proposed by the optical research community to replace SONET/SDH networks [44, 46, 47].

An easy approach to bringing cost effectiveness to optical metro networks is by using the so-called tunable-transmitter fixed-receiver (TT-FR) optical WDM rings. In these settings, every node in the ring is provided with a fixed receiver tuned on a specific (often dedicated) wavelength for the reception of data, together with a tunable transmitter (or a set of fixed transmitters) used to send data on a particular wavelength. Often, the number of wavelengths equals the number of nodes in the ring, thus allowing each node to have a dedicated home channel for reception. Transmission of data to a given node only requires tuning the transmitter to the destination wavelength, since only such a node is expected to be listening on such a wavelength, while all other nodes transparently bypass the information on all wavelengths except its reception one. A large number of TT-FR-based control mechanisms have been proposed in the last decade to manage access to the (shared) dedicated home channel of a given destination node [44].

Thus, given the fact that the access networks of the future are very likely to be optical, and possibly Ethernet-based (EPON and the emerging WDM-EPON), plus the recent efforts put into extending the Ethernet domain toward the metropolitan area, it seems that deploying metro Ethernet (ME) services over transparent all-optical WDM rings could be a possible scenario for the MAN of the future. This chapter focuses on studying the requirements and features ME may demand from a transparent TT-FR-based WDM ring network.

2.2 Problems of Ethernet in the Metro Environment

Metro Ethernet (ME) is the technology for Metropolitan Networks based on the Ethernet standard. It is a cheap and simple extension of Ethernet to a provider network. This would be a natural evolution taking into consideration that Ethernet has become extremely popular in Local Area Networks (LANs). It is more flexible in bandwidth granularity in comparison to SDH. Optical fiber is the main transmission medium for this technology.

Ethernet by itself is not suitable as a core technology as originally defined, because of the following reasons:

- For large networks, MAC forwarding tables also tend to be very large, so this implies a lack of scalability.
- Not very good stability. In comparison to other technologies such as MPLS and SDH, Ethernet has very slow recovery times due to the Spanning Tree Protocols. Its autoconfigurability has demonstrated to be its main flaw as a backbone technology [126].

- Very limited Traffic Engineering features, in addition to unpredictable traffic patterns due to the use of broadcast for learning MAC addresses.

In this sense, there are several desired features which are expected for any carrier-grade technology from the point of view of operators [6, 126, 127]:

1. Scalability: Hundreds of thousands of customers, that is, Metropolitan and Regional areas.
2. Protection: Providers expect 99.999 % of availability in their networks. SDH and SONET are the key to this availability rates thanks to their protection mechanisms: 50 ms link recovery and end-to-end and nodal failure protection mechanisms.
3. Hard Quality of Service (QoS): hard QoS guarantees not only a prioritization of traffic, but also a deterministic performance of the service as well as an enforcement of the QoS parameters across the network. This is key to operators' business model.
4. Service Management: In this sense, providers need to be able to:
   (a) To give service provisioning quickly.
   (b) To monitor the parameters of the given services.
   (c) To troubleshoot problems.
5. Support of TDM: The main issue is to provide this functionality without an exceptional increase of cost.

In a MAN or WAN, regarding connectivity, from the point of view of the customer, there are three kinds of services required to be provisioned according to the definitions made by the Metro Ethernet Forum (MEF) [6–8]. The Metro Ethernet Forum defines certain standards to connect Ethernet LANs across MANs and WANs:

- Point-to-point: E-Lines.
- Point-to-multipoint: E-Trees.
- Multipoint-to-multipoint, called E-LAN.

To provide these services, several technologies have been proposed to improve scalability: Virtual LANs, Provider Bridges, Provider Backbone Bridges and Provider Backbone Bridges - Traffic Engineering which were reviewed in Chapter 1.

2.3 Providing Metro Ethernet on TT-FR WDM Rings

Fig. 2.1 shows a reference scenario to analyse Metro Ethernet over WDM optical rings. In this scenario, three end-stations from Customer Network no. 1 are connected to an IEEE 802.1ah network comprised by four Backbone Edge Bridges (BEB) (each of them interface a given customer network) and four Backbone Core Bridges (BCB). As shown, the four BCBs

(nodes 1, 2, 4 and 5) together with BEB number 3 are aligned along a ring topology. Assuming that the nodes in the ring are connected with optical fibres, then it is possible to build a transparent TT-FR -based optical ring to exploit the benefits of WDM and transparency. To do so, it is first necessary to define a convergence layer that adapts the features of Metro Ethernet to the underlying WDM ring network.

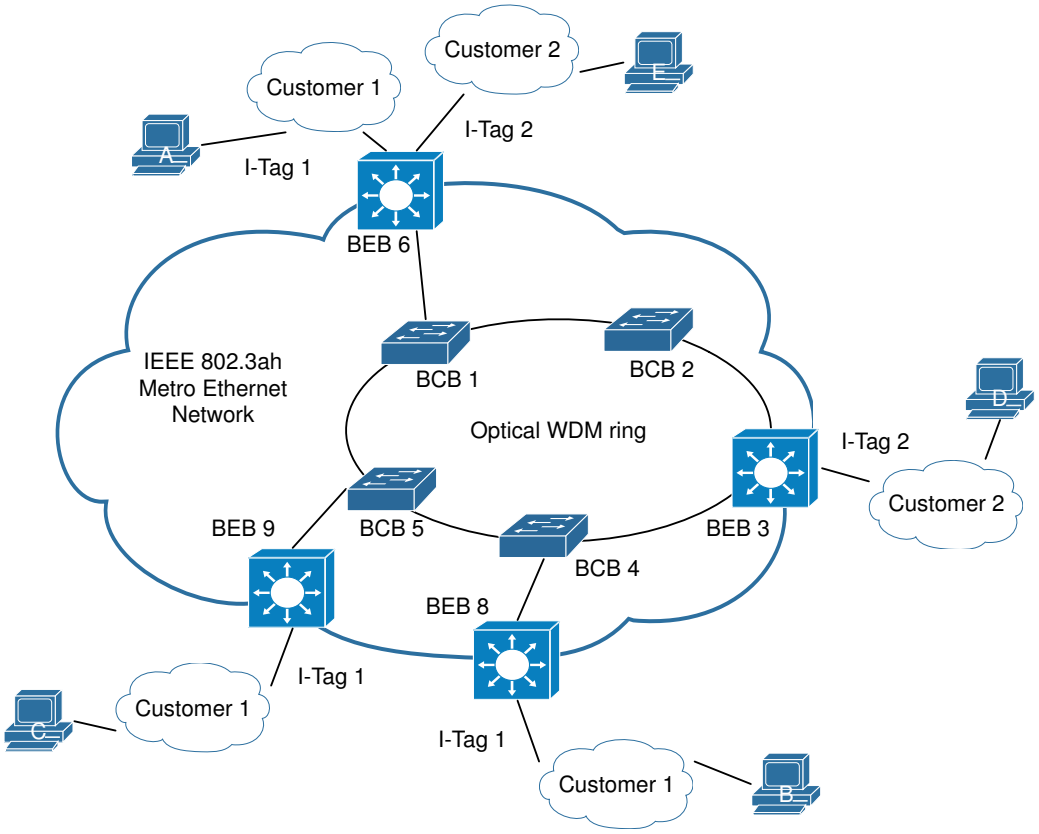


Figure 2.1: Reference scenario

We consider the transmission example of a frame sourced at station A from Customer Network no. 1 and destined to station B. Fig. 2.2 shows the structure of this frame of reference before and after entering the Metro Ethernet network at node BEB 6. Remark that this must traverse nodes: 6 - 1 - 2 - 3 - 4 - 8 to reach its destination.

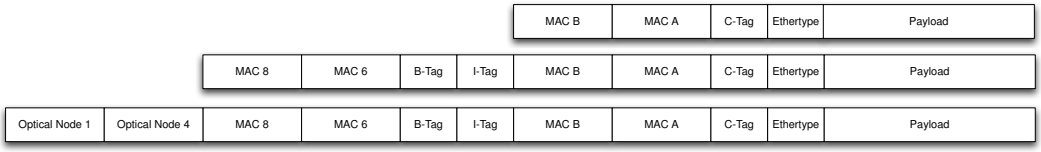


Figure 2.2: Reference frame A: Before entering the ME network (top), after MAC-in-MAC encapsulation (middle), and after entering the Optical Ring (bottom)

If the nodes in the ring are just conventional ME nodes, then the reference frame will suffer Optical to Electrical (OE) conversion, processing, and Electrical to Optical (EO) conversion on every node in the ring, in order to be forwarded to the next one. In this light, the ring is opaque and acts as a store-and-forward network and consequently it does not exploit the benefits of WDM and transparency of optical rings. The next section examines two different ways to provide transparency and exploit the bandwidth provided by WDM: The Logical Full-mesh topology over TT-FR and the Logical Full-mesh topology over TT-FR with hop-by-hop broadcast.

2.3.1 Logical Full-mesh over TT-FR

This case considers a proper TT-FR WDM ring topology (Backbone Bridges 1 to 5 in Fig. 2.1), where every node in the ring is provided with a dedicated home channel for reception (for notation simplicity, node i listens on wavelength λ_i). Essentially, intermediate nodes have the ability to optically bypass traffic from all channels (wavelengths) except its dedicated one, thus it always removes all the packets from its reception home channel. In light of this, the reference frame shown in Fig. 2.2 enters the ME network at BEB 6 and must ideally traverse BCBs 1 - 2 - 3 - 4 to arrive at BEB 8, where the frame egresses the ME network towards its destination station B. The idea is to exploit the potential of WDM and transmit this frame transparently around the optical ring, that is, bypassing (all-optically) nodes 2 and 3. To do so, node BCB 1 must take the following steps upon the reception of our frame of reference: (1) Examine the destination address (this is BEB 8) and understand that the next ME hop is BCB 4; and consequently (2) forward the frame on BCB's 4 home wavelength, that is, λ_4.

To accomplish this, we propose the architecture of Fig. 2.3(a), which combines a conventional Metro Ethernet switch with four in/out ports together with an adaptation box to the TT-FR ring. Four in/out ports interfacing the ME box are necessary to hide the underlying TT-FR network from the upper ME switch. The idea is to *make the Ethernet switch think that it is directly connected to the other four nodes in the ring as in a full-mesh topology*. To do so, it is ideally required that the ME switch sends and receives packets from node i on the i-th port in order to apply backwards MAC learning. It is the role of the adaptation box to translate this behaviour into the proper operation mode of the TT-FR ring.

Basically, when the adaptation box receives a packet from the upper ME switch on input port i, then this packet is intended for the i-th switch in the ring and must be optically transmitted on the home channel of the i-th node, λ_i. Hence, concerning transmission, the adaptation box is in charge of mapping the upper ports to the appropriate wavelengths (frames from port i must go to wavelength λ_i). This way, all-optical transmission transparency is achieved in the WDM ring. There are multiple proposals in the literature to arbitrate the access to a TT-FR ring in order to avoid collisions. In this case, we consider a simple MAC layer that employs a tunable photo-detector (TD) and a Fiber Delay Loop (FDL) following the DBORN architecture [47]. When the station wants to send a frame to a station with home channel λ_i it first checks whether the photo-detector has detected any incoming frame in this wavelength. If it is not the case, it may start transmitting its own frame since, even if some frame arrives just after the check, the FDL will delay it enough time to avoid a collision.

Concerning reception, the operation of the adaptation box must be just the opposite, that

is, mapping the incoming optical frames from the ring to the appropriate port of the ME switch. In other words, frames coming from node i must go to the i-th port. However, a TT-FR node has got no means to distinguish the source node of a given frame received, since all of them arrive on the same wavelength: the node's dedicated home channel. This issue can be solved by the adaptation box by adding an extra optical-ring related header that states which node generated each frame, as well as the destination node. Therefore, every frame entering the optical ring must include an identifier of both the destination and source nodes. This "ring adaptation header" must be added/removed by the adaptation box to the upper ME switch. Then, if the adaptation box is capable of forwarding frames between wavelengths and ME ports appropriately, the ME switch will have a view of the ring as a full mesh topology and backwards MAC learning can operate adequately, as in conventional Ethernet.

Finally, when the ME switch sends a broadcast frame on all its output ports, this frame is then replicated on all the wavelengths of the ring by the adaptation box with the appropriate ring adaptation header: broadcast identifier and source node identifier. Remark that such broadcasting may occur for two reasons in Ethernet: (1) The frame has a multicast or broadcast address (e.g.. ARP queries, IPTV services), and (2) because the MAC destination address is unknown to the switch, and must therefore forward the frame to all output ports.

This strategy is very simple and quite easy to implement but it clearly imposes a capacity penalty since the same frame must traverse the same links on different wavelengths (see Fig. 2.3(b)). The next section proposes an alternative mechanism to reduce such bandwidth consumption in the broadcasting of frames.

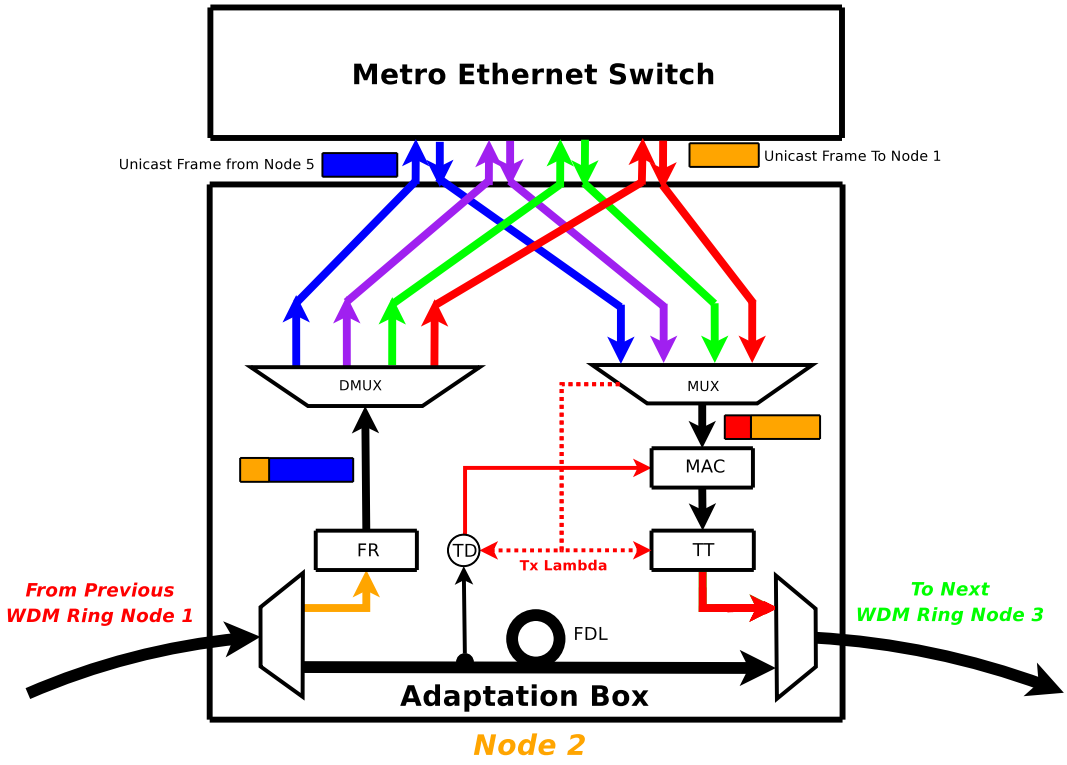
2.3.2 Logical Full-Mesh with hop-by-hop broadcast

This approach combines the benefits of a TT-FR ring concerning optical transparency with the hop-by-hop transmission of broadcast/multicast traffic. Concerning unicast transmission, the adaptation box operates exactly the same as in the previous case, mapping ME ports to wavelengths and including an extra adaptation header to enable the one-to-many reception mapping. The difference with the previous case lies in the transmission of broadcast, multicast and unknown destination frames, which are transmitted hop-by-hop along the ring in order to avoid duplicated frames on all wavelengths (Fig. 2.4(b)).

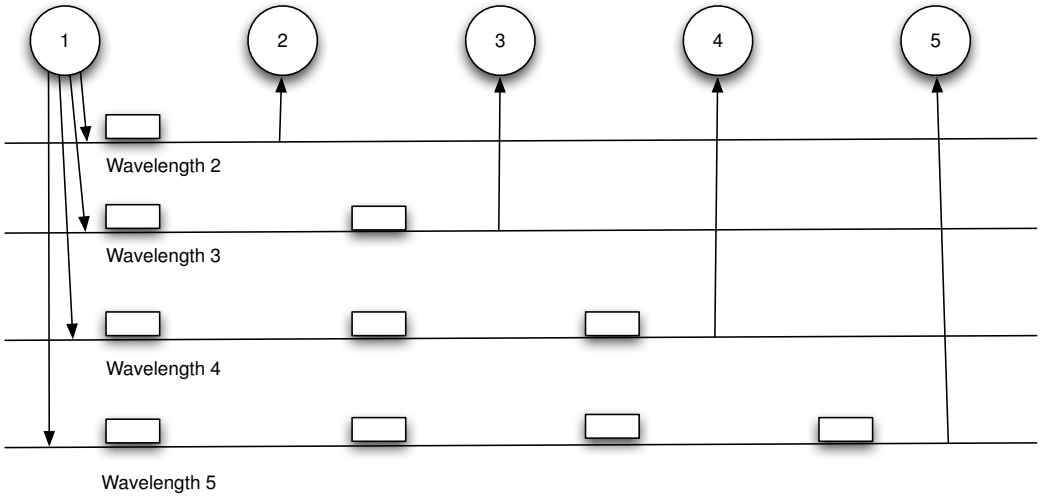
However, since now the broadcast frames must be handled by the Adaptation Box (Fig. 2.4(a)) instead of the Metro Ethernet Switch, it is necessary that the switch sends all broadcast/multicast or frames with unknown destinations by a specific port to the adaptation box instead of duplicating the broadcast frame by all its output ports. Also, the adaptation box needs some additional logic that tells every node whether to forward broadcast frames to the next node in the ring or strip the frame off the ring. The easiest strategy is to employ source stripping, that is, it is the source node which must strip the frame off the ring once this has looped the ring.

Two additional improvements can be done concerning the hop-by-hop broadcasting of frames with unknown destination:

- If a given node acknowledges that this frame is intended for it, then it may strip the frame off the ring, rather than circulating it along the ring.

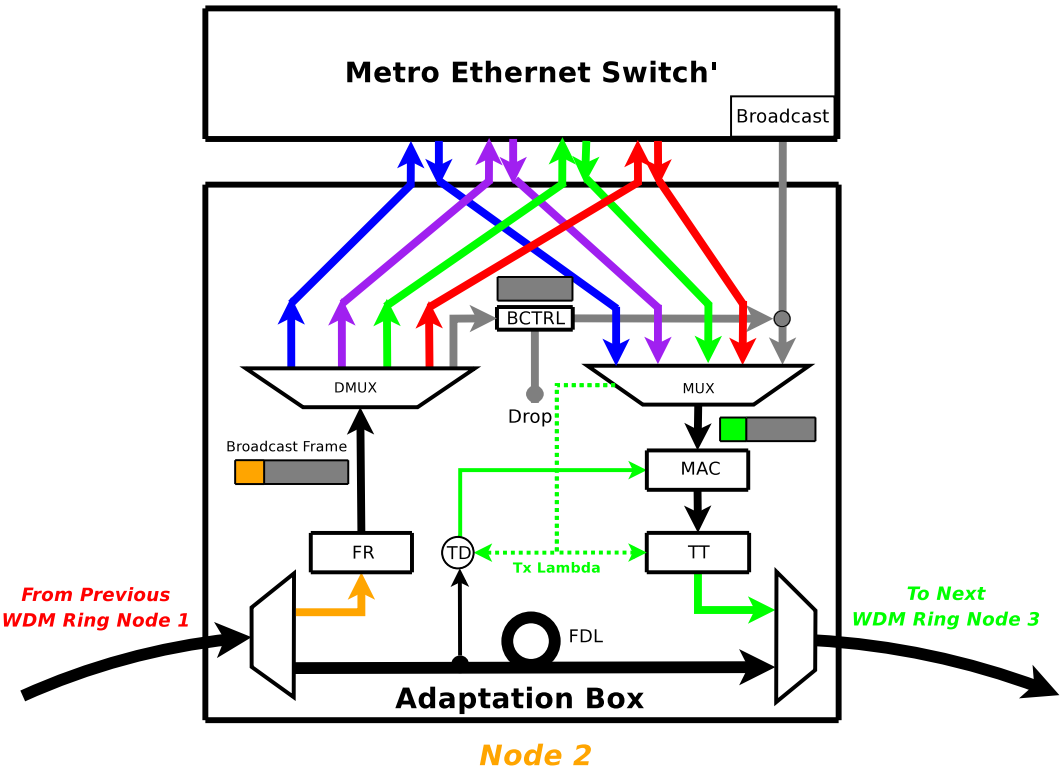


(a) Architecture for a ME TT-FR node

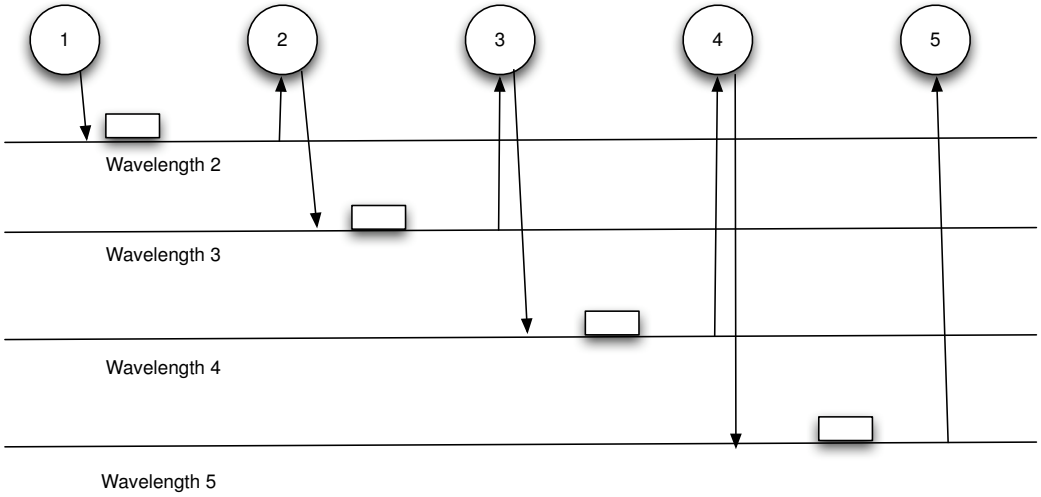


(b) Broadcasting of frames on a Full-Mesh TT-FR WDM ring

Figure 2.3: Architecture for a ME TT-FR node (top); and Broadcasting of frames on a Full-Mesh TT-FR WDM ring (bottom)



(a) Architecture for a ME TT-FR node with hop-by-hop broadcast



(b) Broadcasting of frames on a Full-Mesh TT-FR WDM ring with hop-by-hop broadcast

Figure 2.4: Architecture for a ME TT-FR node with hop-by-hop broadcast (top); and Broadcasting of frames on a Full-Mesh TT-FR WDM ring with hop-by-hop broadcast (bottom)

- If a given node acknowledges that this frame is intended for some other node in the ring, then it may sent it through the dedicated home channel of such destination node, that is, transparently rather than hop-by-hop.

Finally, it is worth noticing that hop-by-hop broadcasting requires all frames to suffer OEO conversion on each node, as it operates in a store-and-forward manner. Additionally, the adaptation box is required to keep internally a sorted list of the next nodes in the ring for every VLAN, since it must know the next hop for every broadcast frame.

2.4 Evaluation

Table 2.1 compares the two ME solutions over TT-FR with conventional ME in terms of average bandwidth consumption and number of OEO conversions per frame, when transmitting both unicast and broadcast frames over a ring topology with N nodes.

Method/Case	Conventional ME	Logical Full-Mesh over TT-FR	Logical Full-Mesh over TT-FR with hop-by-hop broadcast
Unicast: bandwidth consumption	$N/2$	$N/2$	$N/2$
Unicast: OEO conversions	$N/2$	1	1
Broadcast: bandwidth consumption	$N-1$	$(N-1)^2/2$	$N-1$
Broadcast: OEO conversions	$N-1$	$N-1$	$N-1$

Table 2.1: Performance evaluation of ME over TT-FR topologies

Two important conclusions arise from Table 2.1: First, the use of a transparent TT-FR ring reduces the number of OEO conversions per unicast frame to one conversion only. However, the goal of keeping transparent transmission makes frame broadcasting explode in the first solution. To alleviate from this, the hop-by-hop broadcasting of frames reduces bandwidth consumption, but require OEO conversion and processing on each intermediate node. Hence, choosing one strategy or the other depends on the amount of broadcast traffic expected in the network, the number of nodes in the ring, and the traffic load.

2.5 Conclusions

This chapter has reviewed Metro Ethernet switching following the current standards, and has proposed an adaptation layer to deploy ME over transparent WDM optical rings. Basically, such an adaptation layer requires to map out ME switch ports to the appropriate wavelengths of a TT-FR optical ring concerning transmission, and a new adaptation header to differentiate between source nodes concerning reception.

Two solutions have been proposed: The first approach requires no changes on the ME switch but lacks of scalability when broadcast frames comprises an important portion of the total traffic. The second one requires communication between the ME switch and the adaptation box to perform store-and-forward transmission of the broadcast frames in an ordered manner along the ring.

Nevertheless, the use of TT-FR base WDM optical rings results in a virtual all-optical full-mesh topology for unicast frames, since all nodes in the ring are just one-hop distant thanks to its optical transparency nature.

Chapter 3

A hybrid OCDMA-WDM Ring architecture

3.1 Motivation

Optical Code Division Multiple Access (OCDMA) techniques have shown outstanding capabilities in the sharing of optical media, in particular in access networks. However, OCDMA systems may suffer from Multiple Access Interference (MAI) and other kinds of noise when many users access the shared media simultaneously, increasing the BER (Bit Error Rate) to unacceptable levels, that is, a situation at which all combined signals interfere and are lost. This chapter proposes a mixed OCDMA and Tunable Transmitter-Fixed Receiver (TT-FR) WDM and ring architecture at which the ring is split into small-size segments to limit the probability of MAI. Essentially, every segment in the ring has got two hub nodes (on the segment's head and tail) which forwards inter-segment traffic to other hub nodes on dedicated home wavelengths, thus making use of WDM. The access media inside the segment is shared between the nodes by means of OCDMA, and code reuse is possible on different segments. Our performance analysis shows how to split a given ring into segments in order to minimise the BER due to multiple users accessing the network and allow for high bit-rates for a given traffic load. In addition, we analyse the possibility of introducing Forward Error Correction (FEC) at a moderate overhead cost to improve performance.

3.2 The Rendez-vous between WDM and OCDMA

Chapters 1 and 2 already optical ring architectures for Metropolitan Area Networks (MAN), most of today's proposals are based on transparent optical WDM ring networks at which each node in the ring is often offered a dedicated home wavelength for reception, see for instance Hornet, Mawson, RINGO and DBORN [45–47]. Additionally, transmission collisions (when two or more users aim to transmit on the same wavelength of a third node) are avoided by defining Medium Access Control (MAC) protocols that arbitrate channel access, either by using a dedicated control channel, by delaying and inspecting wavelength use [47] or by circulating a token around the ring [44].

Although such TT-FR (Tunable Transmitter-Fixed Receiver) architectures are transpar-

ent and simple to deploy, substantial bandwidth capacity is wasted if nodes have sub-wavelength demands or traffic matrices are very asymmetric, that is, if some nodes have most of the time their reception wavelengths idle, since other nodes do not use them. OCDMA permits bandwidth reuse since all the resources (bandwidth capacity) are shared by all users. There have been some proposals regarding OCDMA-based ring configurations but these are quite limited by MAI because of packet recirculation through the ring [128], leading to self-interference. A possible solution to this issue considers to implement code add/drop multiplexers, but this strategy requires either network synchronisation [129], additional parallel optical sources [130] or taking into account special properties of optical codes [74] which might not be possible for many code families. For this reason, we propose a hybrid OCDMA-WDM segmented ring architecture at which the inter-segment communication follows the same principles of TT-FR WDM rings, while the intra-segment communication occurs in the OCDMA domain. This architecture and its performance evaluation in terms of MAI probability is presented next.

3.3 The hybrid WDM-OCDMA ring architecture

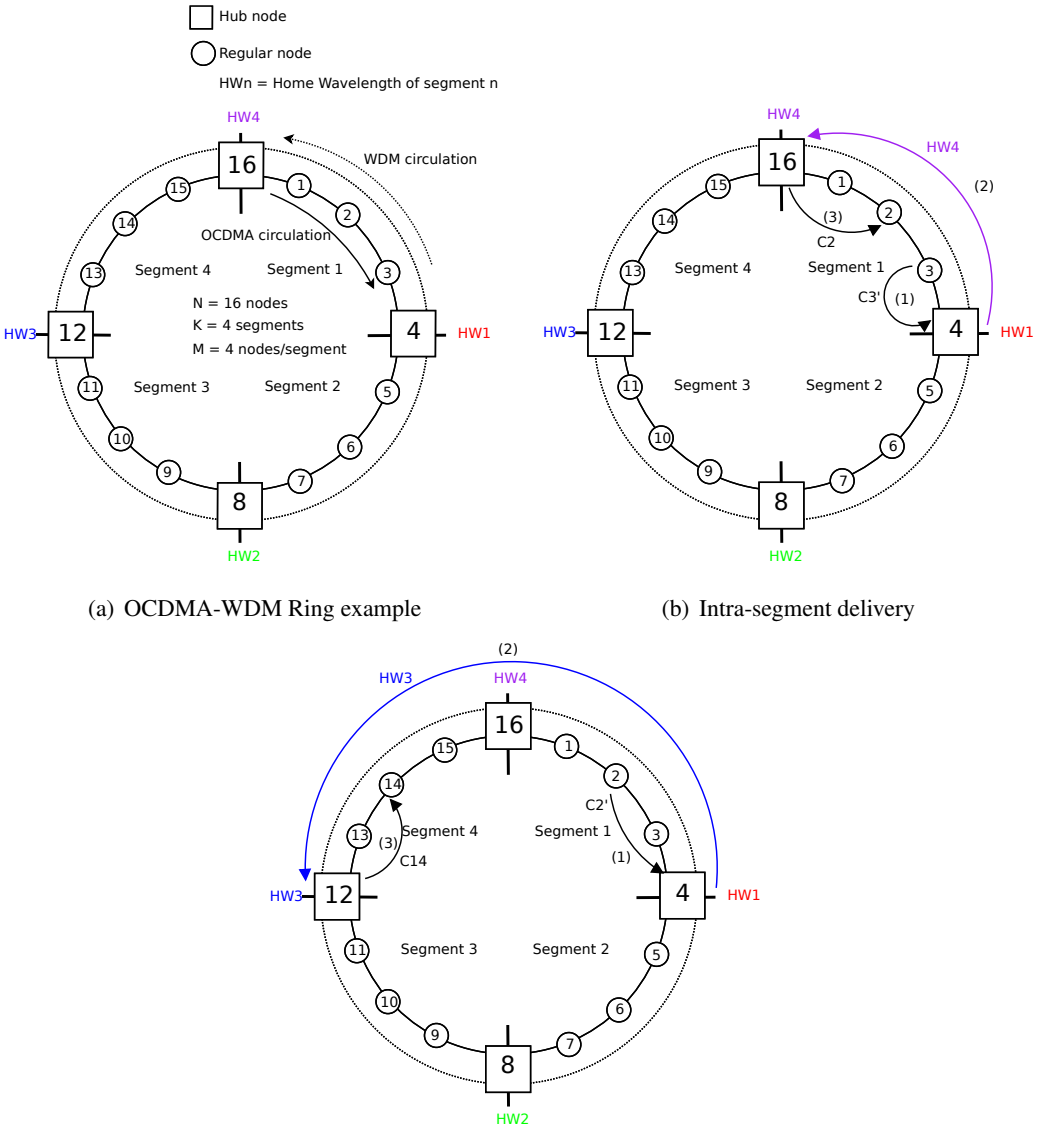
Let us consider the N=16 node unidirectional (clockwise) OCDMA ring of Fig. 3.1(a). This ring is split into a number K of segments ($K = 4$ for the ring of Fig. 3.1(a)) with M nodes per segment. The nodes interfacing adjacent segments, referred to as *hub nodes*, are depicted with squares in the figure, in contrast to *regular* ones depicted with circles. Now, let us consider a single M-sized segment in the ring, see Fig. 3.2.

We assume that each regular node i in the ring has got a unique code for transmission and another one for reception, namely (C'_i, C_i) respectively. Following the testbed of [65], we assume that codes C'_i and C_i use the same codeword but are encoded in orthogonal polarisations, thus not interfering with each other.

That is, every transit node in the segment has got one OCDMA encoder/decoder pair. The hub nodes have got $(M-1)$ encoders/decoder pairs interfacing respectively the next and previous segments they are attached to. The $(M-1)$ encoders of the hub node must match each of the decoders of the regular nodes, and vice versa. This way, the hub nodes have got a means to both collect the traffic from its inbound segment and send traffic to the outbound segment in the ring.

In addition to this, each hub node is provided with at least one home wavelength (HW_n in Fig. 3.1(a)) to receive packets from other hub nodes in the network, as well as one or more Tunable Transmitters that allow them to send packets to other segments in the ring. The number of wavelengths depends on the aggregated traffic generated by each segment. Thus, inter-segment traffic is exchanged between hub nodes in a TT-FR fashion, while intra-segment traffic is exchanged via OCDMA. Both the OCDMA and the WDMA parts have got a dedicated fibre on opposite circulation directions. In Fig. 3.1(a), for instance, the OCDMA circulation uses the clockwise direction while the WDMA direction is counter-clockwise. Finally, we do not assume any time multiplexing, thus each station is able to transmit asynchronously.

As an example of intra-segment packet delivery, consider a packet sent from node 3 to node 2 (Fig. 3.1(b)). To do so, node 3 must encode the packet using its transmission code (C'_3). Then, tail hub node 4 sends the packet to head hub node 16 through HW_4. Finally,



(a) OCDMA-WDM Ring example

(b) Intra-segment delivery

(c) Inter-segment delivery

Figure 3.1: (a) Ring example, (b) Intra-segment delivery and (c) Inter-segment delivery of packets

hub node 16 sends the packet encoded with code C_2, the reception code for node 2. For inter-segment packet delivery, consider a packet sent from node 2 to node 14 (Fig. 3.1(c)). This packet is first encoded at node 2 with code C'_2 and delivered at hub-node 4. Then, hub-node 4 decodes the packet and transmits the packet all-optically to hub-node 12 by tuning its laser on the destination's segment home wavelength, that is, HW_3. Finally, hub node 12 encodes this packet with C_{14} (which is the reception code of node 14) and forwards it locally on its segment towards node 14.

The role of the hub nodes in this architecture is then two-fold:

- First, they strip packets off the ring. Otherwise packets would circulate indefinitely, causing self-interference and noise accumulation. Tail hub-nodes forward only undelivered packets or packets intended for other segments, thus removing already delivered traffic from the ring.

- Second, they partition the ring into small-size segments, limiting the maximum number of OCDMA users simultaneously accessing a given segment to M nodes at most (instead of N). This has a clear impact on reducing the MAI, that might limit the intrasegment performance, as shown in the next section.

Finally, it is worth noticing that code reuse among segments is possible since an OCDMA packet encoded on a given segment is never visible on a different segment. Hence, the hybrid OCDMA-WDM architecture might be possible just with a code cardinality of $(M-1)$ different codewords (M is the segment size). That is, we need $(M-1)$ codewords in one polarisation and $(M-1)$ codewords in the other one. If there were not polarisation multiplexing, we would need a code cardinality of at least $2(M-1)$ codewords for $(M-1)$ nodes.

Concerning hardware requirements, this architecture requires one OCDMA encoder/decoder pair per regular node, and $(M-1)$ encoder/decoder pairs per hub node, that is, $2(M-1)$ encoder/decoder pairs per segment times $K = \frac{N}{M}$, the number of segments, that is:

$$\frac{N}{M} 2(M-1) \approx 2N$$

encoder/decoder pairs for large M. That is, the hardware requirements depend mostly on the ring size, not on how nodes are arranged into segments (segment size). Thus, a number of $(M-1)$ different codewords may coexist on the segment for each polarisation, since packets may be encoded with codes $C_1, \ldots, C_{M-1}$ by the hub nodes in one polarisation or codes $C'_1, \ldots, C'_{M-1}$ by the regular nodes in the other polarisation.

Concerning the dimensioning of the WDM part of the ring, that is, the number of transceivers of the TT-FR ring, each hub node requires:

$$K \cdot \left\lceil \frac{M \cdot R_{b,\text{OCDMA}}}{R_{b,\text{WDM}}} \right\rceil \approx N \frac{R_{b,\text{OCDMA}}}{R_{b,\text{WDM}}}$$

where $R_{b,\text{OCDMA}}$ is the bit rate of the OCDMA segment (for instance, 10 Gbps) and $R_{b,\text{WDM}}$ is the bit rate of a WDM wavelength (for instance, 40 Gbps). Essentially, the hub nodes must collect the total traffic offered by the regular nodes (this is $MR_{b,\text{OCDMA}}$ at most) and forward it to K different segments.

Next, we analyse the MAI probability on a segment of the hybrid OCDMA-WDM architecture.

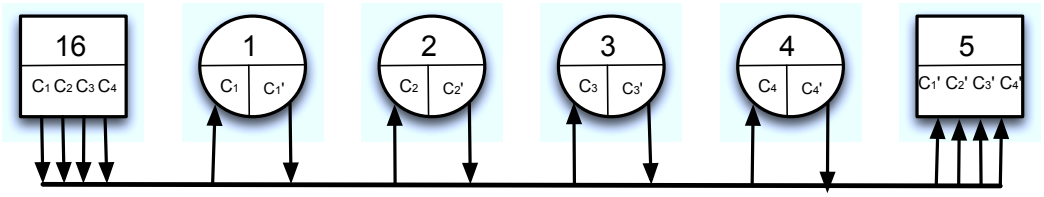


Figure 3.2: OCDMA-WDM segment example (M=5)

3.4 Analysis of the MAI probability

3.4.1 MAI probability for j active users

This section derives the Bit Error Rate (BER) for a hybrid OCDMA-WDM ring with a total number of N nodes and an arbitrary segment size M. Each coding technique has got its own BER equations for a number of j active users. For performance evaluation of the OCDMA-WDM ring, we consider the Spectral Phase Encoding (a coherent OCDMA encoding) technique described in [62] (which was the inspiration for the testbed in [65]) because of its simplicity and feasibility. Nevertheless, it is worth noticing that this analysis may be performed for any other code family.

Now, let A be a random variable that denotes the number of active codes in a given segment at a certain time, thus with sample space $S_A = \{1, \ldots, (M-1)\}$. Additionally, let I refer to the interference random variable which may take two possible values: $I = 1$ if there is MAI, and $I = 0$ otherwise. For the SPE system of [65] the MAI probability for $A = j$ active users is given by [62]:

$$\begin{aligned} P(I=1|A=j) &= \frac{1}{2}\sum_{l=1}^{j-1}\binom{j-1}{l}\left(\frac{1}{2S}\right)^{l}\left(1-\frac{1}{2S}\right)^{j-1-l} \\ &\times \left[1-\gamma^{\beta-1}(l)[\gamma(l)-\rho(l)]\right] \end{aligned} \tag{3.1}$$

where:

$$\gamma(l) = 1 - e^{-\frac{I_{\text{thres}}N_0}{lP_0}} \tag{3.2}$$

$$\rho(l) = 1 - Q\left(\sqrt{\frac{2N_0}{l}}, \sqrt{\frac{2N_0 I_{\text{thres}}}{lP_0}}\right) \tag{3.3}$$

$$\tag{3.4}$$

Here, $Q(a,b)$ refers to the Marcum's Q function:

$$Q(a,b) = \int_b^\infty x \cdot exp\left(\frac{-a^2 - x^2}{2}\right) I_0(ax)dx \tag{3.5}$$

where $I_0(x)$ is the modified Bessel function of the first kind and zeroth order. β is a parameter related to the receiver. It determines that, in order to detect a binary 1, the receiver requires $\beta \cdot \tau_c$ seconds.

We consider very similar parameters to those of the testbed in [65]. We assume $R_b = 10$ Gbps, a pulse width $\tau_c = 450$ femtoseconds, a code cardinality of $N_0 = 64$. Here, T_b is the bit period, then S is computed such that $S = \frac{T_b}{T}$, where T is the width of an encoded pulse. In general, $T_b > T$. We have chosen $T = N_0 \cdot \tau_c = 28.8$ picoseconds. In our case, $S = \frac{T_b}{T} = 3.47$. The bandwidth used is approximately $BW \approx 1/\tau_c$=2.22 THz. Additionally, I_{thres} and P_0 are the threshold current for decision and the received power at the hub node respectively, since it is at the hub node where most interference may occur. Minimum BER is also achieved for ratios $I_{\text{thres}}/P_0 \approx 0.4$, as noted in Fig. 9 of [62]. Finally, we consider that the receivers requires $\beta \cdot \tau_c$ seconds of time to output a bit decision.

3.4.2 Probability of $A = j$ active users in an OCDMA segment

Regular nodes, as well as hub nodes, are assumed to inject traffic to the segment following a Poisson process with an offered traffic of a Erlang:

$$a = \frac{\lambda}{\mu} \tag{3.6}$$

where inter-arrival packet times are exponentially distributed with mean $1/\lambda$, and service times are also exponentially distributed with mean $1/\mu$. Following the Engset's analysis, the probability b of an active source, assuming its offered traffic is a, is given by:

$$b = \frac{a}{1+a} \tag{3.7}$$

Thus, in a given segment, the $(M-1)$ regular nodes inject a Erlang of traffic. Next, we need to derive the amount of traffic injected by the $(M-1)$ encoders of the upstream hub node to each transit node in the segment. Notice that hub nodes do not use OCDMA but WDM to communicate with them. Let a_{ij} denote the amount of traffic offered by any node i to any other node j in the ring ($i, j = 1, \ldots, N$). Thus, assuming that all destination nodes are equally likely, then the total traffic destined to node j in a given segment equals:

$$\sum_{i,i\neq j} a_{ij} = \sum_{i,i\neq j} \frac{a}{N-1} = (N-1)\frac{a}{N-1} = a \tag{3.8}$$

Hence, each code of the hub nodes inject a Erlang of traffic, just like the codes of regular nodes. Thus, on a given segment, we have $2(M-1)$ equal sources of a Erlang, $M-1$ of which are associated to the code space of the regular nodes and the other $M-1$ are associated to the code space of the hub nodes. It should be observed that only $M-1$ of those sources contribute to interference in each orthogonal polarisation, so only $M-1$ are relevant to our analysis. We analyse the probability of interference for any of the two polarisations. Thus,

the probability to have exactly j active sources out of $M-1$ possible on one polarisation is given by:

$$P(A=j) = \binom{M-1}{j} b^j (1-b)^{(M-1)-j} \tag{3.9}$$

where b follows eq. 3.7. Also, the average number of active interfering codes on a segment is $(M-1)b = (M-1)\frac{a}{1+a}$, which increases with the number of nodes per segment and traffic load a Erlang per node. From a design point of view, the goal is to obtain the appropriate segment size M to partition a ring such that the MAI probability, that is the Bit Error Rate (BER) due to interference, on a segment is kept below some value, say for instance 10^{-9}.

3.5 Performance Analysis

From an analysis perspective, we have considered a ring of an arbitrary size N nodes but with different values of M (nodes per segment) and a Erlang of traffic, and studied the resulting segment BER, as:

$$P(I=1) = \sum_{j=0}^{M-1} P(I=1|A=j)P(A=j) \tag{3.10}$$

following the total probability theorem.

In light of this, Fig. 3.3 shows the segment BER for $M = 4$, 8, 16, 32 and 64 nodes for a ring with $N = 64$ nodes. As expected, the segment BER increases with M and a.

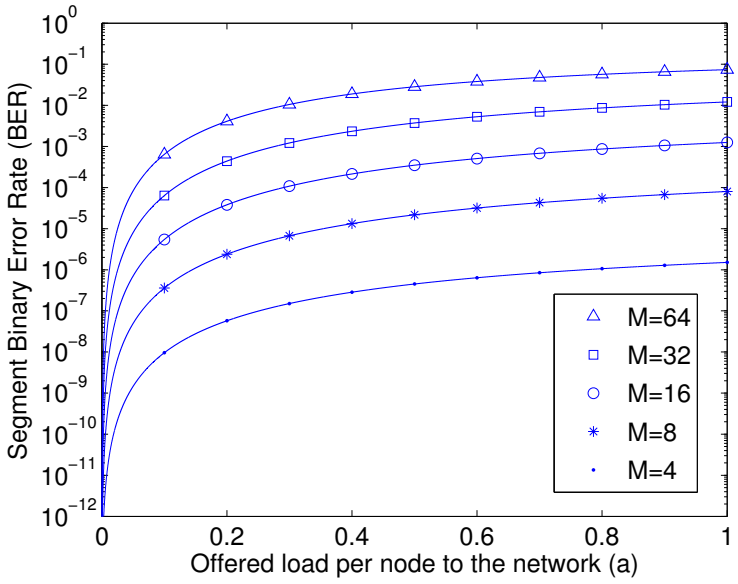


Figure 3.3: Segment BER vs load for $N = 64$ and $M = 4, 8, 16, 32, 64$ for SPE

Fig. 3.3 also shows that doubling the number of nodes per segment implies one order of magnitude more in the BER probability at most traffic loads. That is, the segment BER at

$a = 0.2$ Erlang is about 10^{-6} for $M = 8$ and drops to 10^{-5} for $M = 16$, 10^{-4} for $M = 32$ and 10^{-3} for $M = 64$. In light of these results, the next design question is to study the maximum value of M allowed for a segment size that satisfies a given MAI limit, say for example $P(I = 1) \leq 10^{-9}$, for a given offered load of a Erlang per node.

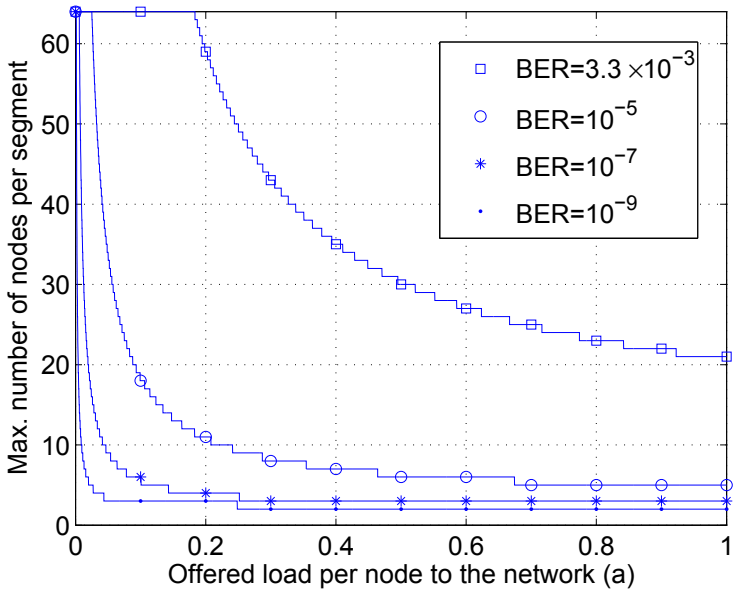


Figure 3.4: Maximum segment size vs load for $N = 64$ for SPE method

In light of this, Fig. 3.4 shows the maximum segment size for different MAI limits ($3.3 \cdot 10^{-3}$, 10^{-5}, 10^{-7}, 10^{-9}), at different traffic loads. For the BER limits of 10^{-9} and 10^{-7}, the maximum segment size quickly decreases and a maximum of $M \leq 2$ or $M \leq 3$ respectively is allowed at network loads of roughly 0.25 Erlang (2.5 Gbps when $R_{b,\text{OCDMA}} =$ 10 Gbps). A significant improvement is observed for a segment BER limit of 10^{-5} since, at the same load $a = 0.25$ Erlang, $M \leq 10$ nodes is supported. The best results are shown for a BER limit of $3.3 \cdot 10^{-3}$ since $M \leq 64$ nodes are supported at $a = 0.183$ Erlang, $M \leq 50$ nodes for a $a = 0.25$ Erlang and $M \leq 21$ nodes at full load: $a = 1$ Erlang (that is, 10 Gbps). It should be noted that, even though the reference level for BER rate is 10^{-9}, it is possible to reduce BER levels of 10^{-5} or even $3.3 \cdot 10^{-3}$ with Forward Error Correction (FEC) techniques at an overhead cost of 7% [131]. For instance, a BER limit of $3.3 \cdot 10^{-3}$ employing 7% FEC overhead is equivalent to a BER limit of 10^{-9} without any FEC.

Finally, concerning spectral efficiency, which we calculate as:

$$\text{Eff} = \frac{(M-1) \cdot a \cdot R_b \cdot (1 - \eta_{FEC})}{BW} \tag{3.11}$$

where $\eta_{FEC} = 0.07$ typically. Table 3.1 shows the spectral efficiency for a BER limit of $3.3 \cdot 10^{-3}$ at medium and full loads.

Table 3.1: Spectral efficiency

BER Target	Load (Erlang)	#Nodes	Eff (bit/s/Hz)
$3.3 \cdot 10^{-3}$	0.3	43	0.1055
$3.3 \cdot 10^{-3}$	1	21	0.167

3.6 Conclusions

This chapter has proposed a novel OCDMA-WDM ring architecture where the ring is split into a number of segments whose bandwidth is shared locally using OCDMA, while inter-segment communication is performed using WDMA. Such a partitioning of the ring into segments permits:

- To limit the maximum number of codes under simultaneous transmission, thus reducing the probability of Multiple Access Interference.
- To reuse code words in other segments, thus reducing the OCDMA code cardinality significantly.

In terms of hardware, a regular node must be equipped with just one OCDMA encoder/decoder pair. Additionally, the hub nodes also require $M - 1$ OCDMA encoders and $M - 1$ OCDMA decoders that is, as many encoders and decoders as nodes per segment for local traffic delivery; plus one or more WDM Tunable Transmitters and Fixed Receivers for inter-segment communication. Additionally, the hub nodes are required to decode and forward all the packets destined to other segments in the ring, which might be a processing burden. The performance analysis shows how to choose the maximum number of nodes per segment M for a target BER probability under different traffic conditions. We have performed the analysis for the Spectral Phase Encoding technique of [62] with the parameters used in the testbed [65]. The results have shown that employing FEC techniques with about 7% overhead may allow segments of up to 50 nodes at low traffic (25%) loads and 21 nodes at full load (100%).

Chapter 4

Buffer Design Under Bursty Traffic with Applications in FCoE Storage Area Networks

This chapter studies the buffer requirements to design the Enhanced Ethernet required by Data Center applications, with bounded buffer overflow probability. The design criteria considers a bursty traffic arrival process, as it shows from conventional read/write SCSI operations in a Fibre Channel over Ethernet (FCoE) scenario. The results show that the buffer must be between one and two orders of magnitude larger than the maximum burst size to provide negligible buffer overflow probability. We complete our study by testing our assumptions with real FCoE measurement traces.

4.1 Introduction and related work

Companies have traditionally run two separate networks in their data center environments: An Ethernet-based LAN for data communications and Internet access with typically 100/1000-Mbps Ethernet NICs, and a Fibre Channel storage area network (SAN) with 1/2/4-Gbps Host Bus Adapters (HBAs) for storage connectivity [102]. Obviously, such a duplicated network architecture requires more equipment and higher Operation, Administration and Maintenance (OAM). Consolidation of the two networks into a single Ethernet-based one, is expected to bring large cost reductions while reducing the OAM. To this end, the IEEE has created the Data Center Bridging (DCB) task group to adapt Ethernet to the high-performance lossless features required in Data Center scenarios. The DCB task group involves several IEEE standards: Priority-based Flow Control (IEEE 802.1Qbb) [103], Enhanced Transmission Selection and Data Center Bridging Exchange (defined in IEEE 802.1Qaz) [104], and Congestion Notification (IEEE 802.1Qau) [105]. A good summary of their role in the definition of such an Enhanced Ethernet can be found in [106].

In particular, the IEEE 802.1Qbb standard (approved in 2011) defines a Priority Flow Control mechanism based on the basic IEEE 802.3x PAUSE semantics. Such a PAUSE frame allows the receiver to communicate buffer availability to the sender for each traffic class on attempts to avoid packet loss due to buffer overflow causes. Essentially, when

the receiver's buffer achieves a certain buffer threshold, it transmits a PAUSE frame to the sender, which stops transmitting data packets. However, the receiver may suffer buffer overflow before such a PAUSE command produces any effect on the sender, especially in cases of large latency between sender and receiver. The authors in [107] provide a worst-case design of buffer thresholds to achieve absolute guarantees of zero packet loss. For instance, a 10-Gbit/s 10-km distant link would require a buffer threshold that leaves 355 KBytes unused (over a total of 480 KBytes). That is, when the receiver buffer achieves 125 KB (i.e. 26% of its total capacity), a PAUSE frame is sent back to the sender. The remaining 355 KB of buffer are left unused in case that, during such an RTT, a data burst of 355 KB is received. Clearly, these numbers are too conservative and have a negative impact on the performance utilisation of the link, especially in long-distance links [107]. This might be an important setback if we are to extend Storage Area Networks (SANs) to longer distances, which has been the object of study for some time [132, 133].

Essentially, FCoE shows a bursty traffic pattern since storage systems structure their data into blocks of several kilobytes. This implies the transmission of several back-to-back packets per read/write operation. However, reserving buffer space for a possible 355KB burst seems rather unreasonable. Nevertheless, failing to notice the bursty nature of FCoE traffic may lead to poor buffer designs that underestimate the buffer-overflow probability, as we show in the following analysis. Thus, the purpose of this work is to analyse the buffer overflow probability of a limited-size queue fed with bursty traffic, similar to that one observed in typical FCoE scenarios. We show that, if the buffer size is correctly designed under the assumption of bursty traffic, the probability to have a buffer overflow is negligible, making the so conservative assumptions of [107] pointless. We validate our assumption both analytically and with a trace captured from several SCSI operations over FCoE [101].

4.2 Analysis

4.2.1 M/M/1/K review

Consider the scenario where packet arrivals follow a Poisson process with rate λ packets/sec. Also, let packet service times be exponentially distributed with mean $E(X) = 1/\mu$, and let $\rho = \lambda E(X)$ refer to the link load. Classical queueing theory states that the buffer overflow probability for such an M/M/1/K queue equals:

$$P_{\text{overflow}} = \frac{1-\rho}{1-\rho^{K+1}}\rho^K, \quad \rho < 1 \tag{4.1}$$

which considers the probability for a random packet arrival to find the queueing system full (in state K).

4.2.2 Analysis of a buffer-limited queue fed with a Burst Poisson Process

Now, consider the case of a queue fed with a Burst Poisson Process (BPP) [134], that is, when packet arrivals occur as bursts. In this case, the burst arrival profile is the same as in the case of a Poisson process, that is, burst inter-arrival times follow an exponential distribution with rate λ_b bursts/sec. However, the difference lies in the fact that every burst is composed of a number of packets i, with $1 \leq i \leq m$, characterised by some discrete

probability distribution. Let q_i refer to the probability that a burst is comprised of exactly i packets. Again, packet service times are assumed to have an exponential distribution with mean $E(X) = 1/\mu$ secs/packet. Thus, a Continuous-Time Markov Chain (CTMC) can be used to characterise the probability to have exactly n packets in the queue, p_n, $n = 0, 1, \ldots, K$. Such probability values are obtained after solving the steady-state equations of the CTMC:

$$pQ = 0 \tag{4.2}$$

$$\sum_{n=0}^{K} p_n = 1 \tag{4.3}$$

where the infinitesimal generator matrix Q follows Eq. 4.4.

$$Q = \begin{bmatrix} -\lambda_b & \lambda_b q_1 & \lambda_b q_2 & \ldots & \lambda_b q_m & 0 & 0 & \ldots & \ldots \\ \mu & -(\lambda_b+\mu) & \lambda_b q_1 & \lambda_b q_2 & \ldots & \lambda_b q_m & 0 & \ldots & \ldots \\ 0 & \mu & -(\lambda_b+\mu) & \lambda_b q_1 & \lambda_b q_2 & \ldots & \lambda_b q_m & 0 & \ldots \\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\ 0 & 0 & \ldots & 0 & \mu & -(\lambda_b+\mu) & \lambda_b q_1 & \lambda_b q_2 & \lambda_b(q_3+\ldots+q_m) \\ 0 & 0 & 0 & \ldots & 0 & \mu & -(\lambda_b+\mu) & \lambda_b q_1 & \lambda_b(q_2+\ldots+q_m) \\ 0 & 0 & 0 & 0 & \ldots & 0 & \mu & -(\lambda_b+\mu) & \lambda_b \end{bmatrix} \tag{4.4}$$

After solving the steady-state probability vector $p = [p_0, p_1, \ldots, p_K]$, the buffer overflow probability must weight the case of i packet arrivals times the probability to find the queue in the n-th state. Let $E(D_K)$ define the average number of dropped packets assuming a burst arrives exactly when the buffer is completely full (that is, in state $n = K$). Then:

$$E(D_K) = \sum_{i=1}^{m} i q_i \tag{4.5}$$

If the buffer is in state $n = K - j$, then the average number of dropped packets is:

$$E(D_{K-j}) = \sum_{i=(j+1)}^{m} (i-j) q_i, \quad j = 0, \ldots, m-1 \tag{4.6}$$

Finally, the buffer overflow probability must be weighted by the probability to find the buffer in each state:

$$P_{\text{overflow}} = \frac{1}{E(D_K)} \sum_{j=0}^{m-1} E(D_{K-j}) p_{K-j} \tag{4.7}$$

It is finally worth noticing that the queue load is given by:

$$\rho = \frac{\lambda_b}{\mu} E(Burst) \tag{4.8}$$

where $E(Burst)$ refers to the average number of packets per burst, computed as:

$$E(Burst) = \sum_{i=1}^{m} iq_i \tag{4.9}$$

The traffic profile observed from different FCoE traces (see Table 4.1) reveals a bursty behaviour that can be characterised by the Burst Poisson Process presented in this section. In particular, the traces show typical burst sizes between 9 and 15 packets of 2KB packets.

4.3 Experiments

4.3.1 Numerical examples

This section examines the buffer overflow probability values at different scenarios, under the assumption that burst sizes are uniformly distributed between 1 and m packets, i.e. $U(1, m)$, or that they have a $\delta(m)$ distribution (that is, burst sizes have a fixed size value).

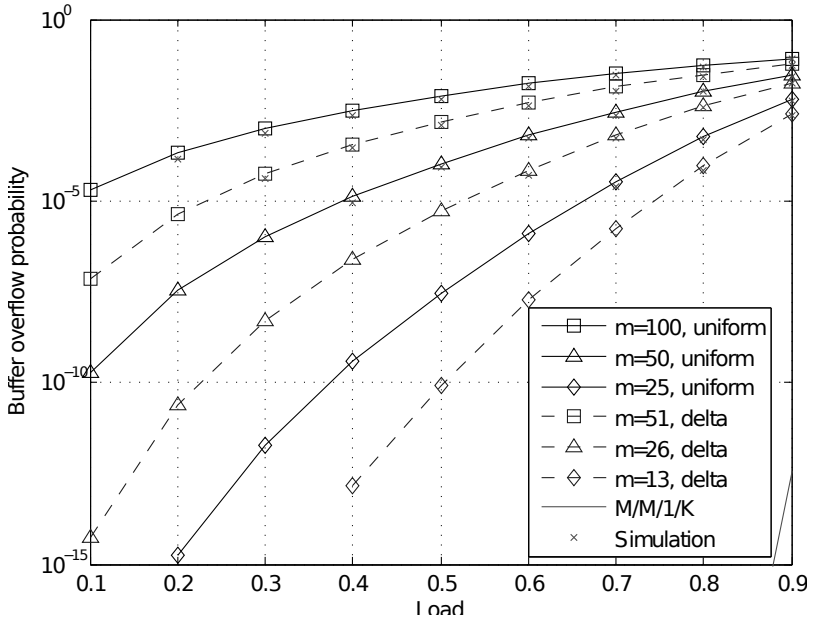


Figure 4.1: Buffer overflow probability, for maximum burst size $m = \{25, 50, 100\}$, $K = 250$ packets.

Fig. 4.1 shows the buffer overflow probability at different load levels and for different values of maximum burst size $m = \{25, 50, 100\}$ for the uniform case and $m = \{13, 26, 51\}$ for the delta case. $K = 250$ packets of buffer size (this is nearly 530 KB for 2172-byte FCoE packets) for both cases. As shown, the buffer overflow probability reaches unacceptable values at high loads especially for large burst sizes. Fig. 4.1 also shows the overflow probability derived from the M/M/1/K model, which demonstrates that such a model cannot be used to design buffers fed with bursty traffic. Only some experiments have been considered due to simulation-time constraints.

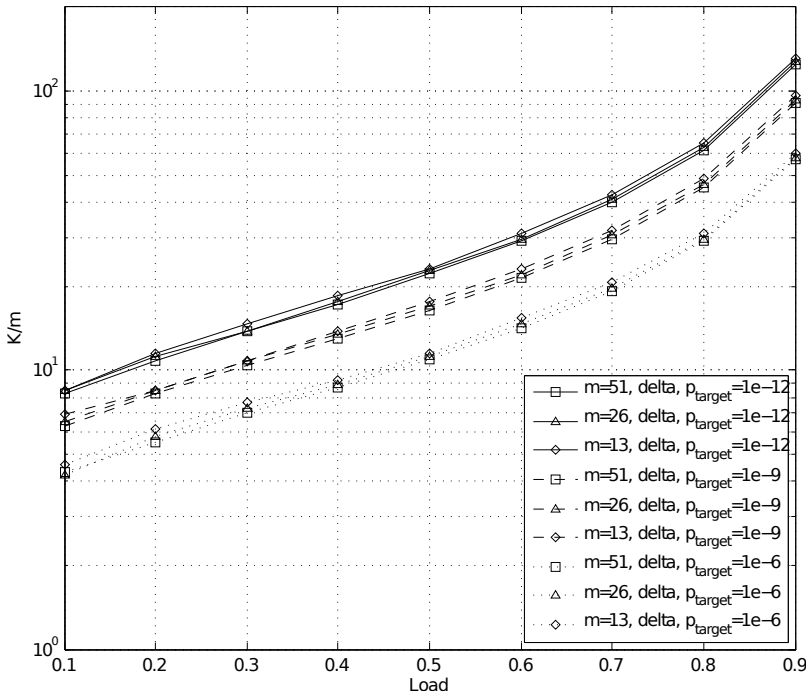


Figure 4.2: Minimum buffer size required to meet $P_{\text{overflow}} < P_{target}$.

The next figure (Fig. 4.2) depicts the minimum buffer size required to guarantee a certain buffer overflow probability target at different load values, following:

$$\text{Find } K \text{ such that } P_{\text{overflow}} < P_{target}$$

for different values of P_{target}, namely 10^{-6}, 10^{-9} and 10^{-12}. The log-y axis depicts the normalised value of K/m, i.e. buffer size in bursts (number of maximum size bursts). As shown, the buffer requirements approximately double per 0.1 increase in terms of load. Furthermore, approximately two orders of magnitude in the K/m ratio are required to guarantee the condition $P_{\text{overflow}} < P_{target}$ at all traffic loads. It is worth noting that, for a given overflow probability, the K/m ratio is almost constant for different burst sizes.

4.3.2 Experiments with traces

This section studies the simulation results with real traces generated with Open-FCoE[1] which is an open-source implementation of the FCoE protocol for the Linux operating system. After setting up the target and the initiator, we have generated seven traces involving different FCoE operations for the experiment. In our scenario, both the FCoE initiator and the target are connected to an Ethernet switch at 1 Gbit/s. One of the switch's ports is configured to replicate all traffic between the target and the initiator towards a high speed packet capture monitoring station, which then captures packet arrival times and sizes. Figs. 4.3(a) and 4.3(b) show the Cumulative Distribution Functions (CDFs) of packet inter-arrival times

[1] http://www.open-fcoe.org/

and burst-size distribution for each experiment of Table 4.1. The inter-arrival times show that most packets (between 75% and 90%) travel back-to-back on the same burst, while the burst-size CDF shows three typical burst sizes: 3 packets (trace 3), 10-11 packets (traces 2, 5 and 7) and 14 packets (traces 1 and 6). The experiments are summarised in Table 4.1.

Exp.	Description	Direction	Average Bitrate (Mbps)	Average burst size (packets)	Burst mode (packets)
1	Copy archive with kernel files	target to initiator	112.42	9.56	14
2	Copy archive with kernel files	initiator to target	223.06	9.24	11
3	Copy kernel files	target to initiator	58.53	3.86	3
4	Copy kernel files	initiator to target	182.84	9.25	11
5	Compile process	initiator to target	43.34	9.23	11
6	Copy source and compiled kernel code	target to initiator	231.81	8.69	14
7	Copy source and compiled kernel code	initiator to target	296.06	9.18	11

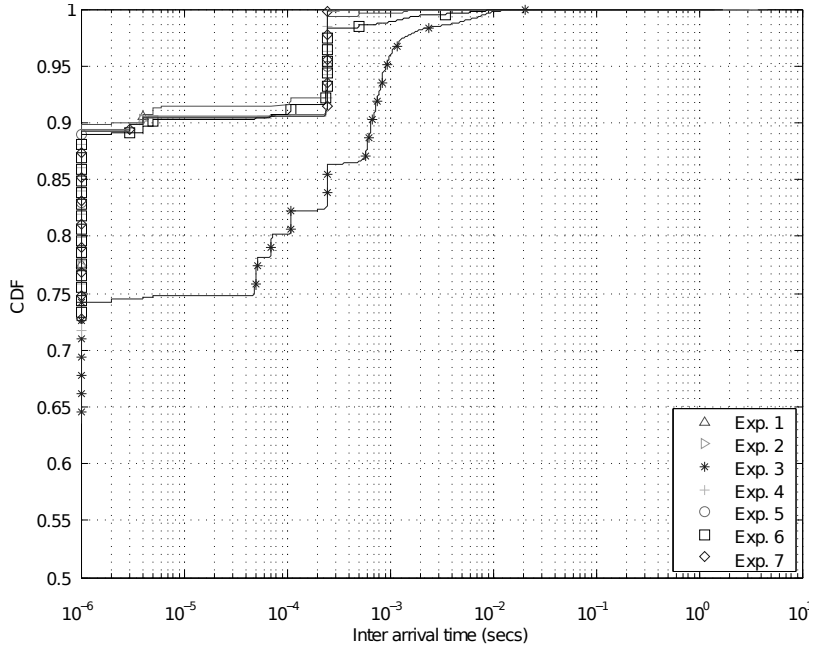
Table 4.1: Experiments

Next experiment (Fig. 4.4) shows the buffer overflow probability whose q_i coefficients are estimated from the experiments of Table 4.1. Two buffer sizes are considered: $K = 100$ (solid) and 250 packets (dashed). Simulations with Poisson arrivals and fixed packet sizes (2172 bytes, to approach a more realistic environment) have been carried with the same q_i coefficients for all experiments, although only those for 1, 3 and 5 are displayed for the sake of legibility. As shown, the BPP model provides an accurate upper bound for the buffer overflow probability, since the simulation considers fixed packet service times rather than exponential ones. Again, only a few simulation points are included due to simulation time constraints.

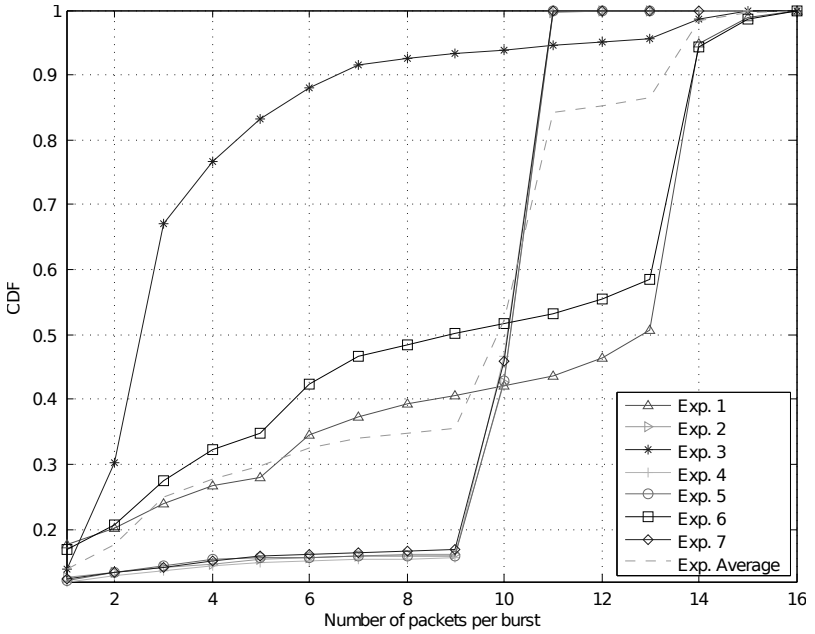
In conclusion, those experiments with highest burst size (experiments 1 and 6) show the largest buffer overflow probabilities, and viceversa. In addition, for a target buffer overflow probability of 10^{-9}, the switch must never exceed 50-55% load for $K = 250$, otherwise there is a high risk of buffer overflow.

4.4 Conclusions

This work studies the buffer size required to have a guaranteed buffer overflow probability value below some target value (typically 10^{-9} or 10^{-12}) for a queue fed with bursty traffic. This analysis has a direct application in designing buffers for Enhanced Ethernet equipment as defined in the Data Center Bridging (DCB) Task group of the IEEE. We show that, to guarantee small buffer overflow probability values, the buffer size must be about one or two orders of magnitude larger than the maximum burst size. We have also tested this assumption with real FCoE traffic measurements and calculated the expected buffer overflow probability for cases with different maximum burst size. The experiments reveal that traffic loads of up to 55% can be considered with relatively small buffers ($K = 250$ packets, i.e. nearly 530KB). In such a case, no flow control is needed since the buffer overflow probability is below 10^{-9}.



(a) Inter-arrival times



(b) Burst size cumulative distribution

Figure 4.3: Traces: (a) Packet inter-arrival times (CDF) and (b) Burst size distribution (CDF)

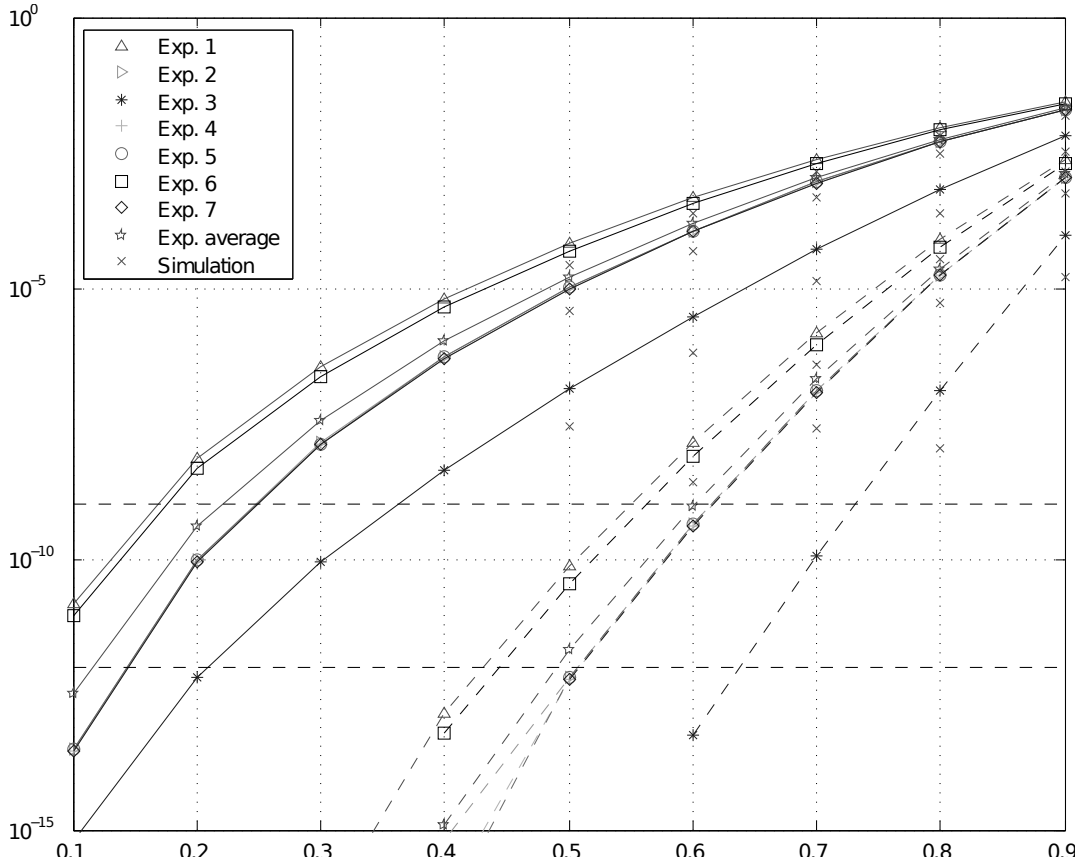


Figure 4.4: Buffer overflow probability of a queue fed with BPP, coefficients obtained from experiments 1-7.

Chapter 5

Packet Coalescing Strategies for Energy Efficiency in the VDE 0885-763-1 Standard for High-Speed Communication over Plastic Optical Fibers

Many recent standards for wire-line communications have included a low-power operation mode for energy efficiency purposes. The recently approved VDE 0885-763-1 Standard for high speed communication over Plastic Optical Fibers has not been an exception. The low-power mode is used when there is no data to be transmitted over the line, thus making consumption more proportional to network load. Furthermore, packet coalescing has been proposed in the literature to minimize the transitions between the low-power and active modes, thus reducing the energy penalties associated to such transitions.

This chapter proposes an adapted version of packet coalescing for the periodic structure of the VDE 0885-763-1 physical layer. Such an algorithm attempts to fulfill active periods of transmission with data, showing an improved energy efficiency over conventional packet coalescing strategies. This conclusion is evaluated both via simulation with synthetic Poisson traffic and with real traces.

5.1 Introduction

In the past few years, several new standards for wireline communications have included energy-efficiency features, mainly on attempts to reduce the overall power consumption and the carbon footprint of communication devices [27]. Probably, the most widely known example is the IEEE 802.3az standard for Energy Efficient Ethernet (EEE) [3, 43], whereby a low-power (also called "sleep") mode was introduced to allow energy savings when no data was pending for transmission. Many subsequent standards have followed this philosophy of introducing a low-power mode that can be used during idle periods of activity. This is also the case of the recently approved VDE 0885-763-1 Standard for High Speed Communication

over Plastic Optical Fibers (POFs) [2, 88].

The use of POFs as a transmission media is interesting due to their immunity to electrical interference, ease of installation and reduced weight and cost [83]. Such positive features make POFs attractive for applications in automotive networks in which weight, cost and interference are important, or for home networking in which installation and cost are critical [83, 86]. Indeed, POFs are already extensively used in multimedia automotive applications [135].

The VDE 0885-763-1 standard defines a configurable physical layer that supports different transmission speeds and link lengths. In this chapter, we focus on the VDE 0885-763-1 configuration for 1 Gb/s bidirectional communication for POF links of up to 50 meters, but it is worth noticing that the standard also allows a 100 Mb/s configuration mode for distances above 50 meters, and even an adaptive rate mode that adjusts the speed to the channel quality conditions. The selection of the 1 Gb/s configuration is driven by the recent creation of a new IEEE 802.3 study group on Gigabit Ethernet over POF [136]. For this standard, the use of the physical layer of the VDE 0885-763-1 standard configured for 1 Gb/s is one of the candidate solutions. This means that the energy efficiency mechanisms studied in this paper may also be used in the future Ethernet standard.

Concerning energy efficiency in Ethernet (the so-called IEEE 802.3az standard), the low-power mode was introduced to allow Ethernet transceivers to save energy when no data was pending for transmission. Indeed, the standard was expected to improve the proportionality between power consumption and network load. However, as demonstrated in [4], the extra energy cost (energy overheads) of entering and exiting this low-power mode has a large energy penalty such that, if transitions occur too frequently, energy savings are reduced.

To minimize the number of low-power to active transitions and viceversa, the use of packet coalescing was further introduced in the literature [42]. The idea of packet coalescing is very simple: once the link has entered the low-power mode, it only transitions back to the active mode when a number of packets (or bytes) are ready for transmission, specified by the max-size parameter. This strategy greatly reduces the number of transitions and their associated energy overheads.

Clearly, implementing packet coalescing has an impact on packet delay, since a packet arrival must wait until a number of other packets have arrived and the max-size criteria is met. To avoid excessively long delays, the max-size threshold criteria is combined with a max-delay threshold which forces the transition to the active mode and subsequent packet departure as soon as sufficient packets have arrived or the waiting delay of the first packet has reached the max-delay limit, whichever occurs first. Thus, fine tuning packet coalescing strategies comprises a trade-off between energy savings and network performance metrics measured in terms of packet delay [137–139].

The use of packet coalescing in the VDE 0885-763-1 standard is analyzed throughout this chapter for first time. Basically, the periodic structure of the physical layer of the VDE 0885-763-1 standard, makes the packet coalescing algorithms used in EEE non optimal as explained in further sections. Therefore, an adapted version of packet coalescing to VDE 0885-763-1 is designed, optimized and evaluated both with synthetic traffic and real packet traces. The results show that packet coalescing can be particularly useful at low loads and especially when packets are short.

Section 1.3.4 reviewed important aspects of both the VDE 0885-763-1 physical layer

and a previous work on energy efficiency, with special focus on Energy Efficient Ethernet was made in Section 1.3.1. Both sections provide useful material to follow the rest of the chapter. The remainder of this work is thus organised as follows: Section 5.1.1 provides an overview of how Energy Efficiency in the VDE 0885-763-1 standard works and its main drawbacks.

Next, section 5.2 introduces a number of packet coalescing strategies for VDE 0885-763-1, which are further evaluated and optimised in a number of scenarios in Section 5.3. Finally, Section 5.4 concludes this work with its main findings and some ideas for future work.

5.1.1 Analysis on Energy Efficiency in VDE 0885-763-1

The VDE 0885-763-1 standard framing and its energy efficiency mechanism was introduced in Section 1.3.4. Although a figure representing the low power idle was introduced in Section 1.3.4, the same figure is reproduced here for convenience purposes. Fig. 5.1 shows the Low Power Idle mode of the VDE 0885-763-1 standard.

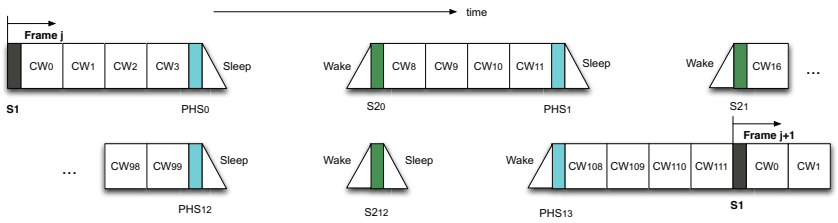


Figure 5.1: Illustration of the use of the low power mode defined in the VDE 0885-763-1 standard

In the VDE 0885-763-1 standard, the energy overhead is mostly due to the transmission of wasted codewords, that is, codewords that carry no user data. For example, if a small 64-byte packet arrives, the link is activated for the whole group, that is, $26.3\mu s$, but only $0.5\mu s$ are actually used in sending such 64 bytes at 1 Gbit/s. In other words, activating a whole group just for the transmission of a short 64-byte packet has the following per-group *efficiency* η of:

$$\eta_{\text{64byte}} = \frac{64}{3290} = 1.95\%$$

The transmission of a large data packet (i.e. 1500 bytes) has a per-group efficiency of:

$$\eta_{\text{1500 byte}} = \frac{1500}{3290} = 46.59\%$$

In order to achieve high efficiency values, it is desirable to achieve near 100% values of such per-group efficiency η, filling as much as possible full groups of four codewords.

The authors in [2] compared the energy consumption vs network load for a link using the VDE 0885-763-1 standard at 1 Gbit/s and a Gigabit Ethernet link under the assumption of Poisson packet arrival times. The simulations in [2] showed that VDE was more efficient than EEE at 1 Gbit/s (see Fig. 5.2, which is Fig. 6 in [2]).

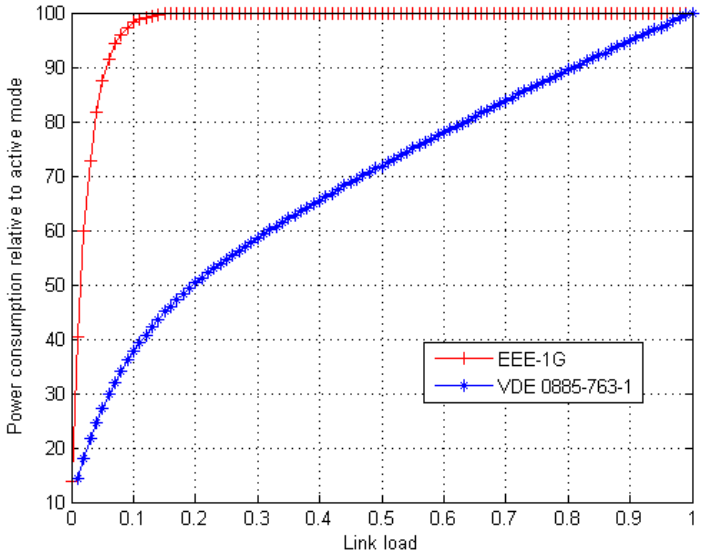


Figure 5.2: Power consumption vs. load for 600-byte packets: comparison of EEE and the VDE 0885-763 standard (from [2])

Essentially, the per-group efficiency values are poor especially at low loads, as demonstrated in the next example.

Example Consider packet arrivals following a Poisson process with rate λ packet/sec and average packet sizes of 600 byte/packet. Let us assume a network load of $\rho = \lambda E(X) = 0.1$, thus:

$$\lambda = \frac{\rho}{E(X)} = 20833.3 \text{ packet/s}$$

since $E(X)$ is the average service time per packet, that is:

$$E(X) = \frac{8 \cdot 600 \text{ bit}}{10^9 \text{ bit/s}} = 4.8\mu s$$

In such a scenario, the number of packet arrivals within an interval of length $T = 26.3168\mu s$ follows the Poisson distribution:

$$P(N(T) = k) = \frac{(\lambda T)^k}{k!} e^{-\lambda T}, \quad k = 0, 1, 2, \ldots$$

When $\rho = 0.1$ (i.e. $\lambda T = \rho \frac{T}{E(X)} = 0.5483$):

$$\begin{aligned} P(N(T) = 0) &= e^{-0.5483} = 0.5780 \\ P(N(T) = 1) &= 0.5483 e^{-0.5483} = 0.3169 \\ P(N(T) = 2) &= \frac{0.5483^2}{2!} e^{-0.5483} = 0.0869 \\ P(N(T) \geq 3) &= 0.0183 \end{aligned} \tag{5.1}$$

which means that about 57.8% of the times, cycles do not carry any traffic and may be switched to the low-power mode, 31.7% of the times a cycle carries a single packet of length $4.8\mu s$ (this results in a per-group efficiency 18.24%), 8.7% of the times the cycles carry 2 packets (per-group efficiency of 36.48%), and finally only 1.8% of the times, the cycle carries three packets or more.

Thus, the average efficiency when a group carries any data (one packet or more) can be approximated by:

$$\begin{aligned} \bar{\eta} &\approx \sum_{k=1}^{\infty} \eta_k p(N(T)=k|N(T)\geq 1) \\ &= \sum_{k=1}^{\infty} \frac{kE(X)}{T} \frac{\frac{(\lambda T)^k}{k!}e^{-\lambda T}}{1-e^{-\lambda T}} \\ &= \frac{\rho}{1-e^{-\lambda T}} \end{aligned} \tag{5.2}$$

This is an approximation since, at high loads, the transmission of the packets in queue may require more than a single cycle. In our example, the average efficiency obtained is: 23.7%.

This is the reason why the energy efficiency mechanism defined for the VDE 0885-763-1 is so inefficient at low loads under Poissonian traffic. Clearly, the way to improve energy efficiency in VDE 0885-763-1 comprises defining mechanims to fill up cycles, that is, increasing the per-cycle efficiency.

5.2 Packet coalescing for the VDE 0885-763-1 standard

5.2.1 Traditional coalescing algorithms

Traditional coalescing algorithms, such as those explained in [42], work as follows: When no data is pending for transmission, the link switches to the low-power mode. However, the link is not put back to the active mode when a single data frame arrives. Instead, *the device waits until a number of s_c bytes have arrived or a timer t_w has expired, whichever occurs first*.

For example, consider the hypothetical case of deterministic 1500-byte packet sizes and a coalescing policy specified by a data-limit value of $s_c = 3000$ bytes and a sufficiently large delay-limit $t_w \to \infty$, such that packet coalescing is always triggered by size, never by delay.

The resulting per-group efficiency of such a packet coalescing algorithm would then be (see Fig. 5.3(a)):

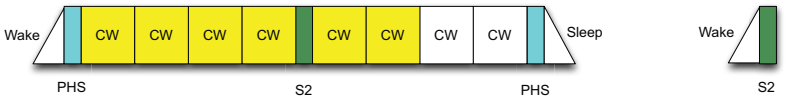
$$\eta = \frac{3000}{3290} = 91.2\%$$

which is close to the 100% target cycle efficiency.

On the other hand, a second packet coalescing algorithm with $s_c = 4500$ bytes is expected to have a poorer average cycle efficiency (see Fig. 5.3(b)):

$$\eta = \frac{1}{2}\left[\frac{3290}{3290} + \frac{4500-3290}{3290}\right] = 68.4\%$$

since this algorithm manages to fill up one group and half of the next group, as observed in the figure.



(b) $b_s = 4500$ bytes

Figure 5.3: Examples of transmission for different coalescing parameters

This behaviour is rather different than that observed in packet coalescing for EEE, where the greater the size threshold, the better efficiency achieved.

Consequently, the overall efficiency of a coalescing algorithm directly depends on its capability to fill groups of codewords entirely, that is, 3290 bytes. In this light, packet coalescing strategies need to be adapted to this behaviour.

5.2.2 Coalescing algorithm proposals for VDE 0885-763-1

This subsection aims at comparing two different strategies to perform packet coalescing for VDE 0885-763-1, namely: *classical* and *strict cycle filling*. We assume the same size and delay thresholds in both of them:

Classical coalescing

A detailed description of this algorithm is shown in Alg. 1. Essentially, in this strategy, the waiting timer starts as soon as the first packet arrival occurs. Next, the algorithm waits until a number of s_c bytes have arrived at the NIC or the maximum waiting-time threshold t_w has expired, whichever occurs first.

When this happens, all packets in the buffer are transmitted at the beginning of the the next cycle. In addition, all packet arrivals during the transmission of these bytes are also transmitted. When the buffer completely empties, the algorithm restarts itself. If new packets arrive during an active cycle after the queue has emptied, these packets have to wait until the s_c and t_w requirements are met again.

This behaviour attempts to imitate the coalescing algorithm defined for Energy Efficient Ethernet, whereby the transceiver is put in the low power mode after transmitting all existing packets in the queue. In this light, it is worth noticing that, although the algorithm defines a maximum transmission size threshold of s_c bytes, in fact more than s_c bytes may be transmitted per active cycle, since more packets may arrive during the transmission of other previously buffered packets. This is best illustrated in the example of Fig. 5.4.

Algorithm 1 Detailed description of the *classic coalescing* strategy for VDE 0885-763-1

```
/* Check if there is an ongoing burst */
if burstongoing == true then
    transmitCurrentPacket();
    if endofburst == true then
        burstongoing = false;
    end if
    waitForNextEvent();
else
    if queue.timeout() == true then
        queue.sendAll();
        burstongoing = true;
        continuetx = true;
    else if queue.bytesInQueue() >= SCBYTES then
        queue.sendAll();
        burstongoing = true;
        continuetx = true;
    else if queue.bytesInQueue() > 0 and continuetx == true then
        queue.sendAll();
        burstongoing = true;
    else
        continuetx = false;
    end if
    waitForNextEvent();
end if
/* continuetx will be set to false too if the transceiver goes to sleep due to inactivity */
```

Algorithm 2 Detailed description of the *strict cycle-filling coalescing* strategy for VDE 0885-763-1

```
/* Check if there is an ongoing burst */
if burstongoing == true then
    transmitCurrentPacket();
    if endofburst == true then
        burstongoing = false;
    end if
    waitForNextEvent();
else
    if queue.timeout() == true then
        if queue.bytesInQueue() >= SCBYTES then
            queue.sendAtLeast(SCBYTES);
        else
            queue.sendAll();
        end if
        burstongoing = true;
        continuetx = true;
    else if queue.bytesInQueue() >= SCBYTES then
        queue.sendAtLeast(SCBYTES);
        burstongoing = true;
        continuetx = true;
    else if queue.bytesInQueue() > 0 and continuetx == true then
        if queue.packetsFitInCycle() == true then
            queue.SendPacketsThatFitInCycle();
            burstongoing = true;
        end if
    end if
    waitForNextEvent();
end if
/* continuetx will be set to false if the transceiver goes to sleep due to inactivity */
```

Strict cycle-filling coalescing

A detailed description of this algorithm can be seen in Alg. 2. In this case, the algorithm behaves similarly to the classical coalescing strategy, but differs in the fact that it transmits only the minimun number of packets to achieve at least s_c bytes *strictly*. Further packet arrivals are only transmitted if they fit in the remaining space of the active cycle. Once such an active cycle has finished, the algorithm is reset.

This algorithm aims at fitting the slotted nature of the VDE 0885-763-1 standard. In this light, the s_c threshold value must be chosen to fill up most of the cycle, that is, close to 3290

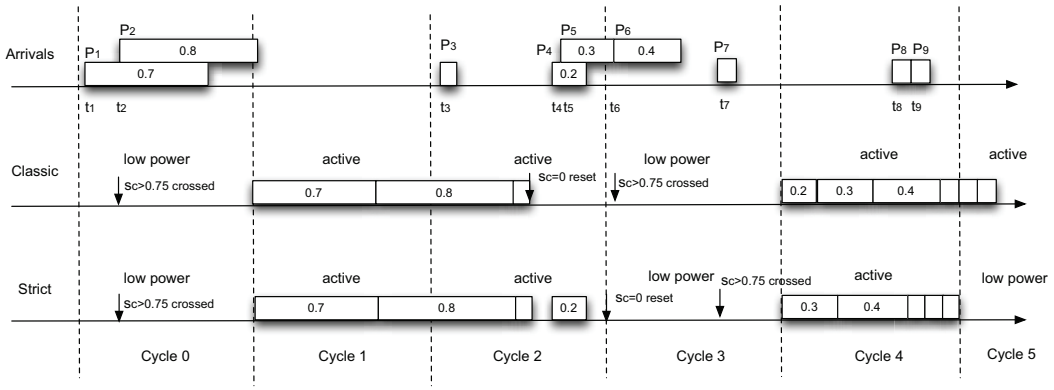


Figure 5.4: Differences between classic and cycle filling packet coalescing strategies, $s_c = 0.75$

Example To illustrate the behaviour of both classic and cycle filling packet coalescing strategies, consider the example of Fig. 5.4, where cycles are normalised to unity length, and packet sizes are smaller than one. We consider both coalescing strategies with configuration $s_c = 0.75$ and $t_W \to \infty$. For simplicity, control packet interval length is considered negligible.

In this example, nine packet arrivals, P_1 through P_9 at times t_1 to t_9 respectively, are considered. Packets P_1 and P_2 arrive at times t_1 and t_2 during a low-power cycle. Because the s_c threshold is crossed, the next cycle becomes active and both packes are transmitted.

P_3, of length 0.1, is inmediately transmitted after P_2 in both algorithms, but for different reasons. The classic algorithm states that any packet arrival before the queue empties must be transmitted. The strict cycle filling algorithm states that new packet arrivals can only be transmitted if they fit in the remaining space of the active cycle; this is the case for P_3.

After the transmission of P_3 the queue empties, and Alg. 1 restarts itself but Alg. 2 does not. For this reason, P_4 can be transmitted during cycle 2 for Alg. 2, but has to wait for Alg. 1 until the s_c condition is met again. On the contrary, P_5 is not transmitted after P_4 since it does not fit in the remaining of cycle 2.

Cycle 3 is a low-power one, but cycle 4 is again active since the s_c thresholds have been crossed during cycle 3.

As shown, the classic coalescing algorithm uses two active cycles (cycles 4 and 5) since new packets (P_8 and P_9) arrive before the queue empties. On the contrary, the strict algo-

rithm manages to fit all packets within one cycle. It is worth remarking than Alg. 2 only transmits more packets than s_c bytes if these fit in the remaining space of an already active cycle.

The average cycle efficiency in both cases are:

$$\begin{aligned}\eta_{Alg.1} &= \frac{1}{4}(0.7+0.8+0.1+0.2+\cdots \\ &\quad +0.3+0.4+0.1+0.1+0.1) = 70.0\% \\ \eta_{Alg.2} &= \frac{1}{3}(0.7+0.8+0.1+0.2+\cdots \\ &\quad +0.3+0.4+0.1+0.1+0.1) = 93.3\%\end{aligned}$$

As shown, the cycle filling case is expected to achieve higher cycle efficiencies than the classic case. Obviously, the example of Fig. 5.4 is not a proper characterisation of the algorithm, since a specific case is being presented, however it illustrates the main differences between the two coalescing algorithms. A statistically relevant evaluation is conducted in the next sections to effectively confirm such intuition. The next section further evaluates both efficiency and packet delay due to coalescing for both algorithms under different network conditions and scenarios.

5.3 Evaluation

5.3.1 Synthetic Poisson traffic: Energy performance

This first set of experiments aims at evaluating the behaviour of the two packet coalescing algorithms (namely, classical and strict cycle-filling) under synthetic Poissonian-like traffic. In this light, we have considered two different packet sizes (short 64-byte packets and long 1500-byte packets), which translate into the following service times at 1 Gbit/s:

$$\begin{aligned}E(X) &= \frac{8\cdot 64 \text{ bit}}{10^9 \text{ bit/s}} = 0.512\mu s \\ E(X) &= \frac{8\cdot 1500 \text{ bit}}{10^9 \text{ bit/s}} = 12\mu s\end{aligned}$$

Packet inter-arrival times are then exponentially-distributed with mean $1/\lambda$, where λ is calculated as:

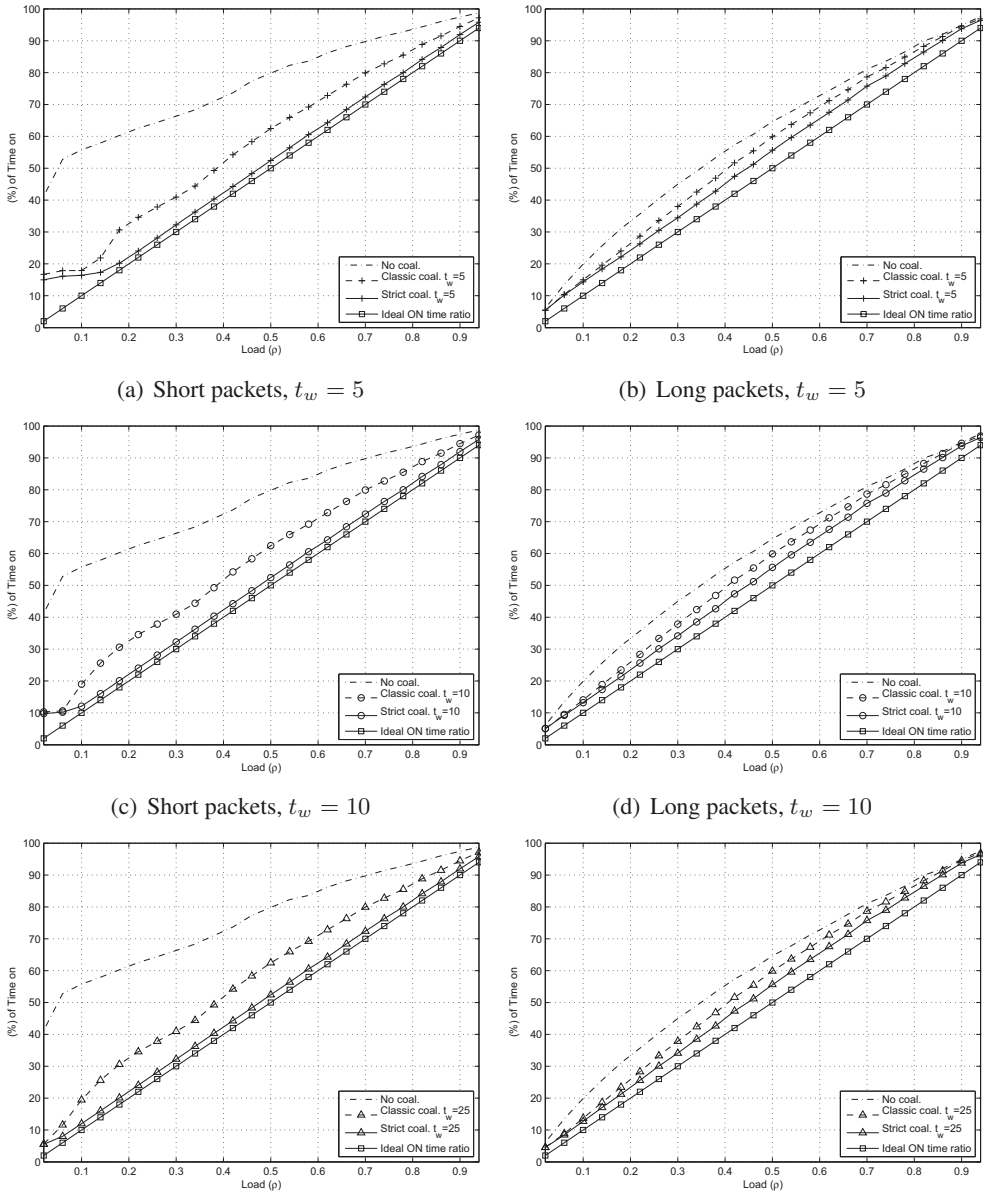
$$\lambda = \frac{\rho}{E(X)}$$

for different load values of $\rho \in (0,1)$.

Fig. 5.5 shows the percentage of cycles that are active cycles (% of time ON) of the two packet coalescing strategies along with the ideal case of proportional consumption-load and the case where no packet coalescing strategy is employed.

In such experiments, the packet coalescing strategies have been configured to follow: $s_c = 3000$ bytes (a bit less than a full cycle) and $t_w = \{5, 10, 25\}$ cycles of maximum waiting time (that is, $161.5\mu s$, $263.2\mu s$ and $657.9\mu s$ respectively, since each cycle lasts for about $26.3\mu s$).

In a nutshell, the results observed reveal that:



(a) Short packets, $t_w = 5$

(b) Long packets, $t_w = 5$

(c) Short packets, $t_w = 10$

(d) Long packets, $t_w = 10$

(e) Short packets, $t_w = 25$

(f) Long packets, $t_w = 25$

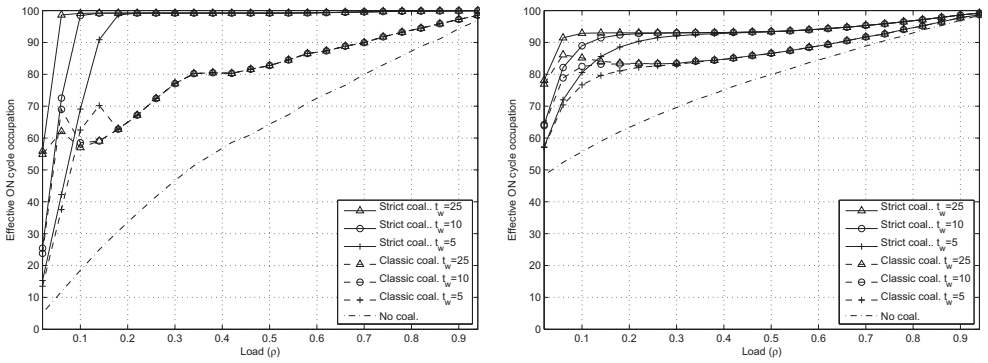
Figure 5.5: Percentage of time active versus load for different packet lengths and t_w

1. Employing any coalescing algorithm clearly outperforms energy-efficiency without coalescing, especially when packets are short.

2. The t_w parameter has a moderate impact on the energy consumption results. Only at low loads large values of t_w improve the energy-consumption figures. However, such small improvement at low-loads is at the expense of larger average delay values, as shown in the next section.

The strict cycle-filling coalescing algorithm is closer to the ideal proportional consumption-load straight line than the classical coalescing algorithm in all cases. When packet sizes are large, the difference between the algorithms and the ideal linear case is not significantly important, but still the strict cycle-filling algorithm outperforms over the rest. Indeed, the strict cycle-filling manages to transmit all packets using a smaller number of active cycles, since it fills cycles better leading to a higher cycle efficiency.

This is clearly observed in Figs. 5.6, where the average cycle efficiency is depicted. As shown, the cycle filling coalescing algorithm approaches 100% cycle utilisation when packets are short, and about 90% when packets are long, at loads greater than 20%, i.e. $\rho > 0.2$.



(a) Short packets (b) Long packets

Figure 5.6: Average cycle efficiency with different packet coalescing strategies

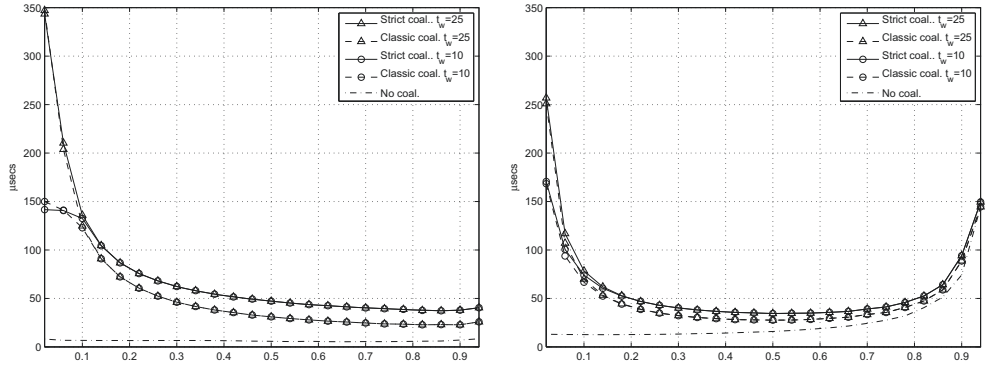
5.3.2 Synthetic Poisson traffic: Delay analysis

Finally, Fig. 5.7 shows the average delay experienced by packets under the two strategies, at different traffic loads ρ and for different values of t_w, i.e. 10 and 25 cycles respectively, that is, $263.17\mu s$ and $657.92\mu s$. The size threshold is fixed to $s_c = 3000$ bytes.

Essentially, both coalescing strategies impose a delay penalty over the case where no-coalescing is employed. However, in most cases, the strict cycle-filling strategy penalises delay more than the classical coalescing. The reason behind is that the classical coalescing algorithm has a number of cases where packets are transmitted earlier than in the cycle filling case, although this behaviour is the opposite in a few cases. This feature allows the classical coalescing case to perform better in terms of delay, but worse in terms of efficiency.

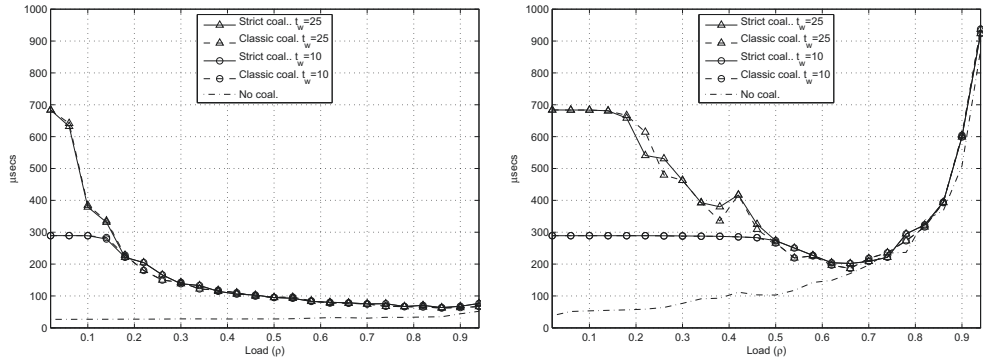
In addition to the average delay, Fig. 5.8 shows the maximum delay observed by packets under the two coalescing strategies for cases with short and long packets. As shown,

the maximum delay is mostly influenced by the maximum waiting time parameter of the algorithm, that is t_w. At medium and high loads, the effect of t_w is negligible, since the algorithms are mostly triggered by the size constraint s_c.



(a) Short packets, $t_w = 10, 25$ (b) Long packets, $t_w = 10, 25$

Figure 5.7: Average packet delay with different packet coalescing strategies ($s_c = 3000$ bytes fixed)



(a) Short packets, $t_w = 10, 25$ (b) Long packets, $t_w = 10, 25$

Figure 5.8: Max packet delay with different packet coalescing strategies ($s_c = 3000$ bytes fixed)

5.3.3 Experiments with real-traces at 1 Gbit/s

This section evaluates the performance of the coalescing algorithms with real 1 Gbit/s traces collected from a number of different scenarios. The traces collected contain packet arrival and packet service times, thus allowing to estimate the behaviour of the different coalescing strategies. Different performance metrics have been computed for each monitored trace, as shown next.

The first scenario under study considers a residential user under two typical cases: (1) Youtube video streaming; (2) BitTorrent file sharing. This user is connected by a 1 Gbit/s Ethernet link to his/her HFC access (50/5 Mbit/s asymetric down/upstream). The second scenario considers a University access link (1 Gbit/s capacity) with highly multiplexed Internet

traffic from students and staff members. Finally, the third scenario considers a number of server traces collected from Google's data centres. The traces comprise three typical server types: a file server which is also involved in search queries, a second server only devoted to search queries, and a their one which acts as both file and application server. All these servers are connected via Gigabit Ethernet.

Table 5.1 shows the cycle efficiency values η obtained under the two coalescing strategies, with different t_w parameters. Table 5.2 shows the average delay obtained for the two coalescing algorithms.

				$t_w = 5$		$t_w = 10$		$t_w = 25$	
Scenario	Direction	Link Load	No coal	Classic	CycFil	Classic	CycFil	Classic	CycFil
BitTorrent	Input	0.61	18.39	21.58	22.50	24.95	25.41	30.84	31.64
BitTorrent	Output	1.75	36.61	41.30	41.85	49.56	50.12	60.17	60.99
Youtube	Input	0.0067	2.33	2.35	2.42	2.55	2.55	2.87	2.92
Youtube	Output	0.21	12.50	13.08	13.32	15.16	15.48	16.18	16.73
University	Input	10.79	40.73	59.80	65.14	69.23	75.44	75.94	82.58
University	Output	17.43	49.31	73.24	79.09	75.71	81.62	76.50	82.39
Data Center 1	Input	1.21	5.78	13.08	14.21	20.92	22.08	41.25	42.75
Data Center 1	Output	52.15	92.77	93.28	93.58	93.80	94.21	94.49	94.95
Data Center 2	Input	8.52	58.64	64.99	65.71	69.71	70.28	76.96	77.64
Data Center 2	Output	7.25	45.78	54.45	55.60	59.85	60.72	68.11	68.72
Data Center 3	Input	0.65	5.32	8.26	8.84	11.23	11.71	19.00	19.42
Data Center 3	Output	4.03	30.19	37.84	39.08	43.98	44.91	55.97	56.53

Table 5.1: Experiments with traces. Link load (%) and average cycle efficiency (%) for the classic and cycle filling algorithms with different t_w values.

				$t_w = 5$		$t_w = 10$		$t_w = 25$	
Scenario	Direction	Link Load	No coal	Classic	CycFil	Classic	CycFil	Classic	CycFil
BitTorrent	Input	0.61	12.32	106.99	104.91	181.86	182.92	386.08	387.89
BitTorrent	Output	1.75	11.98	117.60	116.89	183.92	184.01	324.90	326.98
Youtube	Input	0.0067	11.75	140.37	123.98	231.28	230.34	476.33	467.73
Youtube	Output	0.21	13.95	77.24	81.63	92.59	104.08	121.27	156.60
University	Input	10.79	12.91	69.41	73.31	99.54	106.53	154.21	163.32
University	Output	17.43	12.10	58.21	62.79	68.45	73.21	77.28	81.76
Data Center 1	Input	1.21	7.56	87.67	80.47	153.51	146.29	349.23	346.47
Data Center 1	Output	52.15	137.82	140.37	144.16	141.05	146.31	142.87	148.48
Data Center 2	Input	8.52	61.71	101.05	101.76	127.59	129.51	205.34	209.72
Data Center 2	Output	7.25	224.24	286.78	283.35	335.87	332.41	475.28	474.15
Data Center 3	Input	0.65	10.21	102.70	95.01	180.73	173.48	398.13	392.13
Data Center 3	Output	4.03	171.72	254.91	250.29	324.57	318.89	508.14	503.95

Table 5.2: Experiments with traces. Link load (%), average delays ($\mu seconds$) for the classic and cycle filling algorithms with different t_w values.

As observed, employing any coalescing strategy significantly improves the cycle efficiency in all experiments. In particular, the strict-cycle filling performs better than the classical coalescing strategy especially at medium to high loads and when packet sizes are small. In addition, the value of t_w plays a critical role in both energy savings and average delay experienced by packets.

In the first scenario, for the residential user, we observe that the link load is very small

(up to 2% in the downlink direction, bittorrent case). In such a case, important energy savings are achieved only when using a significantly large value of t_w, as observed in the last column. However, in such cases, the average delay due to coalescing is also significantly increased.

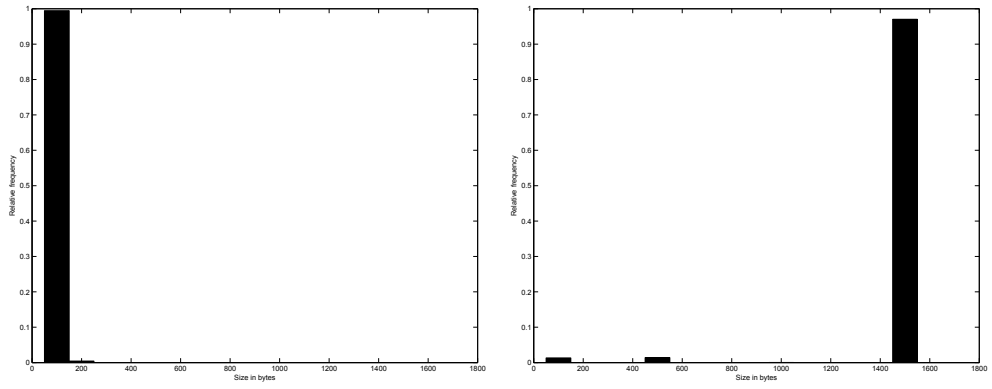
In the second scenario, the university access link, the difference in energy performance between the strategies are significantly large. As observed, the cycle efficiency goes from 40.73% without coalescing, up to 82.58% for the strict cycle-filling coalescing with $t_w = 25$ in the input case, and from 49.31% to 82.39% in the output case. This translates into large energy savings.

Finally, in the third scenario, the cycle efficiency improvements differ from one data set to another.

However, the behaviour observed is consistent with the previous findings: (1) Strict cycle-filling outperforms over classical coalescing; (2) The value of t_w plays a critical role in the cycle filling efficiencies, (3) especially at low loads and (4) for long packets.

Concerning average coalescing delay, the results are consistent with those observed from the synthetic Poisson traffic traces. Essentially, the two algorithms show very similar average delays; in some cases, the classic algorithm shows smaller delays and viceversa. We believe that the average packet size plays an important role in the differences between the two algorithms.

Take for instance the example of the Data Center 1 trace. The packet size histograms are depicted in Fig. 5.9. As shown, packets are small in the input direction but large in the output direction. The average delay in the input direction is smaller for the strict cycle-filling algorithm than for the classic algorithm. The reason for this has to do with the fact that the strict cycle-filling algorithm allows to add new packet arrivals after the queue has emptied only if they fit in the remaining space of the already active cycle. This is obviously easier when packets are small, as it is the case in the input direction.



(a) Histogram 1: Data Center 1, Input

(b) Histogram 2: Data Center 1, Output

Figure 5.9: Packet size histograms for Data Center, Trace 1

5.4 Summary and discussion

This work has proposed two packet coalescing strategies to further improve energy efficiency in high-speed communications over Plastic Optical Fibers following the VDE 0885-763-1 standard. Both strategies, namely classical and strict cycle-filling attempt to fill up the active $26\mu s$ cycles defined as the minimum active periods for data transmission. Each strategy has its pros and cons: while the cycle filling strategy priorises the improvement of cycle efficiency, in some cases, it imposes more delay to individual packets than the classical coalescing algorithm.

These conclusions have been validated with both synthetic Poisson traffic and real traffic traces collected from a number of 1 Gbit/s scenarios. Indeed, the experiments have revealed that the strict cycle-filling coalescing algorithm shows significantly better energy performance, especially at low loads and when the average packet size is large. Furthermore, the s_c and t_w parameters have been shown to clearly impact the two performance metrics of the coalescing algorithms, namely energy efficiency and coalescing delay.

Future studies may involve profiling different scenarios and investigating the influence of the coalescing algorithms on different traffic patterns in terms of energy efficiency and coalescing delay. In addition, future work shall study how to dynamically adapt the s_c and t_w parameters of the coalescing algorithm to changes in the traffic conditions, mainly load, average size and inter-arrival times of packets.

Chapter 6

Conclusions and future work

6.1 Summary and Conclusions

In this chapter, the main conclusions of this thesis's contributions are outlined as well as the future lines of work for those lines of research.

The first contribution (chapter 2) studies how to provide Metro Ethernet services over transparent tunable-transmitter fixed-receiver WDM optical ring networks. A new adaptation layer of ME to WDM, called "Adaptation Box", is proposed, and its benefits and drawbacks are studied. Basically, such an adaptation layer requires to map ME switch ports to the appropriate wavelengths of a TT-FR optical ring concerning transmission, and a new adaptation header to differentiate between source nodes concerning reception. It is shown that such a transparent WDM ring network can be seen as a logical full-mesh topology by the upper ME layer, thus reducing to one the number of optical-electronic-optical conversions per unicast frame. Additionally, two different approaches are proposed in the case of broadcast/multicast traffic, since this may bring scalability difficulties in ring topologies. Two solutions have been proposed concerning frame broadcast. The first approach, called Logical Full Mesh, requires no changes on the ME switch but lacks scalability in broadcast traffic delivery. The second one, called Logical Full Mesh with Hop by Hop Broadcast, requires additional coordination between the ME switch and the adaptation box to perform store-and-forward transmission of broadcast frames in an ordered manner along the ring. This allows to save bandwidth by forwarding a frame in a hop by hop basis, using only one wavelength to forward the frame to the next destination in each case instead of occupying several wavelengths simultaneously to send the same frame.

Obviously, certain delay would be added, but this delay is less important than the bandwidth savings that would be obtained by using this broadcast technique. It also would require specific tables at the edge nodes to guarantee that the frame is forwarded properly along the ring using the least possible bandwidth and OEO conversions. That would require the frame to be forwarded in such a way that it is passed to the nearest node in each hope to improve bandwidth efficiency.

In all cases, the use of TT-FR-based WDM optical rings results in an all-optical virtual full mesh topology for unicast frames, since all nodes in the ring are just one hop distant thanks to its optical transparent nature.

Further study on Metropolitan networks is made in the second contribution (chapter 3)

in the context of Optical Division Multiple Access. Optical Code Division Multiple Access (OCDMA) techniques have shown outstanding capabilities in the sharing of optical media, in particular in access networks. However, OCDMA systems may suffer from Multiple Access Interference (MAI) and other kinds of noise when many users access the shared media simultaneously, increasing the BER (Bit Error Rate) to unacceptable levels. That is, a situation at which all combined signals interfere and are lost. This contribution proposes a mixed OCDMA and Tunable Transmitter-Fixed Receiver (TT-FR) WDM and ring architecture at which the ring is split into small size segments to limit the probability of MAI. Essentially, every segment in the ring has got two hub nodes (on the segment's head and tail) which forwards inter-segment traffic to other hub nodes on dedicated home wavelengths, thus making use of WDM. The access media inside the segment is shared between the nodes by means of OCDMA, and code reuse is possible on different segments. Our performance analysis shows how to split a given ring into segments in order to minimise the BER due to multiple users accessing the network and allow for high bit-rates for a given traffic load. In addition, we analyse the possibility of introducing Forward Error Correction (FEC) at a moderate overhead cost to improve performance.

The network architecture suggested in the second contribution by partitioning an OCDMA ring in segments permits:

- To limit the maximum number of codes under simultaneous transmission, thus reducing the probability of Multiple Access Interference.
- To reuse codewords in other segments, thus reducing the OCDMA code cardinality significantly.

In terms of hardware, a regular node must be equipped with just one OCDMA encoder/decoder pair. Additionally, the hub nodes also require $M-1$ OCDMA encoders and $M-1$ OCDMA decoders that is, as many encoders and decoders as nodes per segment for local traffic delivery; plus one or more WDM Tunable Transmitters and Fixed Receivers for inter-segment communication. Additionally, the hub nodes are required to decode and forward all the packets destined to other segments in the ring, which might be a processing burden. The performance analysis shows how to choose the maximum number of nodes per segment M for a target BER probability under different traffic conditions. We have performed the analysis for the Spectral Phase Encoding technique with the parameters used in the testbed of reference for that model. The results have shown that employing FEC techniques with about 7% overhead may allow segments of up to 50 nodes at low traffic (25%) loads and 21 nodes at full load (100%).

The third contribution (chapter 4) studies the buffer requirements to design the Enhanced Ethernet required by Data Center applications, with bounded buffer overflow probability. The design criteria considers a bursty traffic arrival process, as observed from conventional read/write SCSI operations in a Fibre Channel over Ethernet (FCoE) scenario.

Specifically, the buffer size required to have a guaranteed buffer overflow probability value below some target value (typically 10^{-9} or 10^{-12}) for a queue fed with bursty traffic is studied. This analysis has a direct application in designing buffers for Enhanced Ethernet equipment as defined in the Data Center Bridging (DCB) Task Group of the IEEE. We show that, to guarantee small buffer overflow probability values, the buffer size must be about one

or two orders of magnitude larger than the maximum burst size. Interestingly enough, our analysis shows that there seems to be an almost constant ratio between the maximum buffer capacity and the maximum burst size for a given block probability and load value.

Regarding Energy Efficiency, many recent standards for wire-line communications have included a low-power operation mode for energy efficiency purposes. The recently approved VDE 0885-763-1 Standard for high speed communication over Plastic Optical Fibers has not been an exception. The low-power mode is used when there is no data to be transmitted over the line, thus making consumption more proportional to network load. Furthermore, packet coalescing has been proposed in the literature to minimize the transitions between the low-power and active modes, thus reducing the energy penalties associated with such transitions. The fourth contribution (chapter 5) proposes an adapted version of packet coalescing for the periodic structure of the VDE 0885-763-1 physical layer. Such an algorithm attempts to fulfill active periods of transmission with data, showing an improved energy efficiency over conventional packet coalescing strategies.

In this sense, the fourth contribution has proposed two packet coalescing strategies to further improve energy efficiency in high-speed communications over Plastic Optical Fibers following the VDE 0885-763-1 standard. Both strategies, namely classical and strict cycle-filling attempt to fill up the active $26\mu s$ cycles defined as the minimum active periods for data transmission. It has been shown that, for synthetic traffic, the strict coalescing algorithm achieves much better results for both short and long packets even for low waiting times. These results have been compared with those of a previous paper that considered no coalescing, providing a point of comparison in similar conditions.

Each strategy has its pros and cons: while the cycle filling strategy priorises the improvement of cycle efficiency, in some cases, it imposes more delay to individual packets than the classical coalescing algorithm. In the fourth contribution, the average and maximum delays have been computed to provide an idea of the trade offs associated with the proposed algorithms. High delays are achieved at low loads, mainly due to the maximum waiting times. At high loads, by contrast, the delay increases because packets are buffered for a long time before they are processed. This affects way more long packets than short packets, because the transmission time of long packets has an obvious effect in waiting time that short packets do not have.

These conclusions have been validated with both synthetic Poisson traffic and real traffic traces collected from a number of 1 Gbit/s scenarios. Indeed, the experiments have revealed that the strict cycle-filling coalescing algorithm shows significantly better energy performance, especially at low loads and when the average packet size is large. Furthermore, the s_c (minimum assembly size) and t_w (maximum waiting time) parameters have been shown to clearly impact the two performance metrics of the coalescing algorithms, namely energy efficiency and coalescing delay. The simulations with traces show that the energy savings depend very much on the traffic pattern.

6.2 Future work

Future lines of work for the first contribution include further simulation of the architecture proposed by using a discrete event simulator, under both Poisson and self-similar traffic. Further validation can be achieved by using real traces obtained from real network-

ing scenarios to test the robustness of the suggested Metro Ethernet ring architecture of the first contribution. Further validation of the architecture with realistic metro scenarios can be done in future contributions. Examples of possible scenarios are:

- Multicast video distribution (eg: IPTV).
- VPN service for enterprise site interconnection.

The multicast video distribution, either for videoconference purposes or simply for IP TV distribution is a key scenario of applicability, given the evident advantages of the proposed architecture for multicasting. In this type of scenario, there would be a central node that connects the ring with the higher levels of the operator's network. From this point the video traffic, would be multicasted to the rest of the nodes in the ring. In the enterprise site interconnection scenario, an evaluation of the behaviour of the architecture for typical unicast traffic could be provided. Other scenarios worth of investigation are multicast videoconference and, to achieve synergies with the third contribution, to evaluate the integration of WDM rings with SANs. As mentioned earlier in the state of the art, several research contributions have tried to integrate WDM and SAN networks. By testing the architecture in the first contribution to provide SAN interconnection in a metropolitan ring may open interesting possibilities. In both scenarios, performance metrics such as delay, jitter and packet losses can be studied.

Further lines of work for the second contribution may include trying different OCDMA families and compare results. Furthermore, more realistic results might be achieved by using physical level simulators with different traffic patterns to calculate the Multiple Access Interference in a more precise way. This may be achieved by means of Poisson and self-similar traffic or even by using, for example, realistic Ethernet traces. Additionally, new architectures can be tested and its performance compared with that of the architecture proposed in the second contribution. Furthermore, the same future scenarios suggested for the first contribution could be applied for this case, since our main target for the second contribution was also a MAN/WAN scenario.

Future work for the third contribution will consider simulation studies by using different traffic patterns and constructing new models for them. Other lines of work can try to investigate the design of the upper and lower thresholds for Priority Flow Control under self-similar traffic considering several distances. By designing these thresholds with criteria that take into account the statistical properties of the traffic involved, it is reasonable to assume that adequate levels of those thresholds may be achieved that guarantee maximum traffic throughput with minimal overflow probability.

Such a study would involve investigating how likely, given a certain traffic pattern, it is to reach a buffer overflow (or underflow) from a given buffer occupancy level for several load values. A first approach could be to use self-similar traffic for different load levels and distances. These results can be compared with those of a scenario with PFC disabled and thus evaluate the trade offs between simplicity, throughput, delay and packet loss probability due to buffer overflow.

Nevertheless, static thresholds might not be the best answer. One possible line of future work may involve estimating such thresholds with an heuristic algorithm. This algorithm could monitor the current load in a link and adjust the thresholds with a certain periodicity

to adapt to the current network situation. This would ensure safer operation at high loads and looser threshold values for low load situations. Another interesting research line in this field might include guaranteeing capacity fair share among end devices at switches. All these lines of future work for the third contribution can be generalised for other applications of Convergence Enhanced Ethernet, not only SANs.

Future studies for the fourth contribution may involve profiling different scenarios and investigating the influence of the coalescing algorithms on different traffic patterns in terms of energy efficiency and coalescing delay. In addition, future work shall propose new algorithms to dynamically adapt the s_c and t_w parameters of the coalescing algorithm to changes in the traffic conditions, mainly load, average size and inter-arrival times of packets. It should also be noted that the fourth contribution mainly studied the 1 Gbps case, but other binary rates are allowed by the VDE 0885-763-1 standard. Future studies can evaluate the differences in Energy efficiency among the different binary rates.

Further studies should consider different use cases and choose a traffic model that best suits the needs of the POF networking environment evaluated, since POF might be used for either home networking or internal vehicle data transmission, which probably have very different requirements. Traces from vehicle data transmission may provide a good insight on the effectiveness of our packet coalescing algorithm, since the main tests performed for them have been taken from traditional data environments and the traffic pattern in a car may be very different. Another interesting line of work could be to advance in the mathematical modelling of the system described in the fourth contribution to provide an analytical insight and design by means of, among others, Markovian models.

Finally, there can be some integration research, especially between the areas of the third contribution (Enhanced Ethernet) and the fourth contribution (Energy Efficiency in POF), to provide energy savings and evaluating the trade offs in performance, namely delay, loss ratio, etc., in a context in which high performance and reliability are a major requirement due to the needs of storage technologies of a dependable transmission level to avoid loss or corruption of information. In this sense, there have been standardisation efforts in the field of energy efficiency in Fibre Channel over Ethernet. Studying the viability of FCoE in a POF environment with energy efficiency is a really interesting possibility.

In this sense, one possible future contribution could study the Energy Efficiency of FCoE in POF with and without PFC. A first approach could be to apply mathematical modelling as mentioned above to see the differences between enabling and disabling PFC. Our current POF simulator can be extended to support a PFC mechanism in the presence of Poisson and self similar traffic.

6.3 List of publications related to this thesis

6.3.1 Main thesis publications

- C1: Rodríguez de los Santos, G.; Urueña, M.; Hernández, J.A.; Larrabeiti, D., "On providing metro ethernet services over transparent WDM optical rings," *Network, IEEE*, vol.25, no.1, pp.14-19, January-February 2011

- C2: Rodríguez de los Santos, G.; Hernández, J.A.; Urueña, M.; Seoane, I.; Larrabeiti,

D., “Study of a hybrid OCDMA-WDM segmented ring for metropolitan area networks,” *High Performance Switching and Routing (HPSR), 2011 IEEE 12th International Conference on*, pp.83-88, 4-6 July 2011

- C3: Rodríguez de los Santos, G.; Urueña, M.; Muñoz, A.; Hernández, J.A., “Buffer Design Under Bursty Traffic with Applications in FCoE Storage Area Networks,” *Communications Letters, IEEE*, vol.17, no.2, pp.413-416, February 2013

- C4: Rodríguez de los Santos, G.; Reviriego, P.; Hernández, J.A., “Packet Coalescing Strategies for Energy Efficiency in the VDE 0885-763-1 Standard for High-Speed Communication over Plastic Optical Fibers”, *Submitted to the IEEE/OSA Journal of Optical Communications and Networking.*

6.3.2 Other publications

- Seoane, I., Rodríguez de los Santos, G., Hernández, J. A., Urueña, M., Romeral, R., Cuevas, A., Larrabeiti, D.,Â “Analysis of delay mean and variance of collision-free WDM rings with segment recirculation of blocked traffic”, *Photonic Network Communications*, Junio 2011

- Urueña, M., Muñoz A., Aparicio R., Rodríguez de los Santos, G., “Digital Wiretap Warrant: Protecting civil liberties in ETSI Lawful Interception”. *Submitted to Transactions on Information and System Security*.

- Rodríguez de los Santos, G., Hernández, J.A., Urueña, M., Muñoz A., “A Bloom Filter-based monitoring station for a Lawful Interception Platform”. *Multimedia Communications, Services & Security (MCSS) 2014*, 2014. **Best paper award**

- Aparicio, R., Urueña, M., Muñoz A., Rodriguez de los Santos, G., Morcuende S., “INDECT Lawful Interception platform: Overview of ILIP decoding and analysis station”. *Jornadas de Ingeniera Telematica (JITEL) 2013*, October 2013.

- Rodríguez de los Santos, G., Hernández, J.A., “Analysis and modeling of multipath resilient ad-hoc network”, *BONE Summer School 2009*, September 28th-29th 2009 Krakow, Poland.

- Rodríguez de los Santos, G., Romeral, R., Larrabeiti, D., Velasco, L., Agraz, F. Spadaro, S., “Emulated MPLS-ASON/GMPLS Inter-connection test-bed”, *VIII Workshop in G/MPLS Networks*, June 29th 2009, Girona, Spain.

- Rodríguez de los Santos, G., Larrabeiti, D., Seoane, I., “Soporte Multitrayecto en red ad-hoc basada en una propuesta de extensión de AODV”, *XVIII Jornadas Telecom I+D*, October 29th-31st 2008, Bilbao, Spain.

- Rodríguez de los Santos, G., Seoane, I., Larrabeiti, D., “MPLS-based local protection in ad-hoc networks”, *15th Eunice International Workshop (poster)*, September 2009, Barcelona, Spain.

References

[1] U. Troppens, M.-H. Wolfgang, R. Wolafka, E. Rainer, and N. Haustein, *Storage Networks Explained*. Chichester, West Sussex, The United Kingdom: John Wiley and Sons, 2009.

[2] P. Reviriego, P. Pérez de Aranda, and C. Pardo, "Introducing energy efficiency in VDE 0885-763-1 standard for high speed communication over plastic optical fibers," *IEEE Communications Magazine*, vol. 51, no. 8, pp. 97–102, Aug. 2013.

[3] *IEEE 802.3az standard, Energy Efficient Ethernet*, 2010.

[4] P. Reviriego, J. A. Hernández, D. Larrabeiti, and J. A. Maestro, "Performance evaluation of Energy Efficient Ethernet," *IEEE Communications Letters*, vol. 13, no. 9, pp. 697–699, Sept. 2009.

[5] "IEEE Standard for Information technology– Local and metropolitan area networks– Specific requirements– Part 3: CSMA/CD Access Method and Physical Layer Specifications Amendment 6: Physical Layer and Management Parameters for Serial 40 Gb/s Ethernet Operation Over Single Mode Fiber," *IEEE Std 802.3bg-2011 (Amendment to IEEE Std 802.3-2008)*, pp. 1–53, March 2011.

[6] R. Sanchez, L. Raptis, and K. Vaxevanakis, "Ethernet as a carrier grade technology: developments and innovations," *Communications Magazine, IEEE*, vol. 46, no. 9, pp. 88–94, September 2008.

[7] "MEF 6.1: Ethernet Services Definitions - Phase 2," *MEF 6.1: Ethernet Services Definitions - Phase 2*, pp. 1–59, April 2008.

[8] "MEF 10.1: Ethernet Services Attributes Phase 2," *MEF 10.1: Ethernet Services Attributes Phase 2*, pp. 1–65, October 2009.

[9] K. Kompella and Y. Rekhter, "Virtual Private LAN Service (VPLS) Using BGP for Auto-Discovery and Signaling," *IETF RFC 4761*, January 2007.

[10] V. Kompella and M. Lasserre, "Virtual Private LAN Service (VPLS) Using Label Distribution Protocol (LDP) Signaling," *IETF RFC 4762*, January 2007.

[11] E. Rosen and Y. Rekhter, "BGP/MPLS IP Virtual Private Networks (VPNs)," *IETF RFC 4364*, February 2006.

[12] "IEEE Standards for Local and Metropolitan Area Networks: Media Access Control (MAC) Bridges," *IEEE Std 802.1D-1990*, pp. 1–, 1991.

[13] "IEEE Standard for Local and Metropolitan Area Networks - Common Specification. Part 3: Media Access Control (MAC) Bridges - Amendment 2: Rapid Reconfiguration," *IEEE Std 802.1w-2001*, pp. i–, 2001.

[14] "IEEE Standards for Local and Metropolitan Area Networks— Virtual Bridged Local Area Networks— Amendment 3: Multiple Spanning Trees," *IEEE Std 802.1s-2002 (Amendment to IEEE Std 802.1Q, 1998 Edition)*, pp. 1–211, 2002.

[15] "IEEE Standard for Local and metropolitan area networks–Media Access Control (MAC) Bridges and Virtual Bridged Local Area Networks–Amendment 20: Shortest Path Bridging," *IEEE Std 802.1aq-2012 (Amendment to IEEE Std 802.1Q-2011 as amended by IEEE Std 802.1Qbe-2011, IEEE Std 802.1Qbc-2011, IEEE Std 802.1Qbb-2011, IEEE Std 802.1Qaz-2011, and IEEE Std 802.1Qbf-2011)*, pp. 1–340, June 2012.

[16] R. Perlman, D. Eastlake, D. Dutt, S. Gai, and A. Ghanwani, "Routing Bridges (RBridges): Base Protocol Specification," *IETF RFC 6325*, July 2011.

[17] "IEEE Standard for Information Technology- Telecommunications and Information Exchange Between Systems- Local and Metropolitan Area Networks- Common Specifications Part 3: Media Access Control (MAC) Bridges," *ANSI/IEEE Std 802.1D, 1998 Edition*, pp. i–355, 1998.

[18] "IEEE Standards for Local and Metropolitan Area Networks: Virtual Bridged Local Area Networks," *IEEE Std 802.1Q-1998*, pp. i–, 1999.

[19] "IEEE Standard for Local and Metropolitan Area Networks Virtual Bridged Local Area Networks," *IEEE Std 802.1Q-2005 (Incorporates IEEE Std 802.1Q1998, IEEE Std 802.1u-2001, IEEE Std 802.1v-2001, and IEEE Std 802.1s-2002)*, pp. 1–285, 2006.

[20] S. Blake, D. Black, M. Carlson, E. Davies, Z. Wang, and W. Weiss, "An Architecture for Differentiated Services." *IETF RFC 2475*, December 1998.

[21] "IEEE Standard for Local and Metropolitan Area Networks—Virtual Bridged Local Area Networks—Amendment 4: Provider Bridges," *IEEE Std 802.1ad-2005 (Amendment to IEEE Std 802.1Q-2005)*, pp. 1–74, May 2006.

[22] "IEEE Standard for Local and metropolitan area networks – Virtual Bridged Local Area Networks Amendment 7: Provider Backbone Bridges," *IEEE Std 802.1ah-2008 (Amendment to IEEE Std 802.1Q-2005)*, pp. 1–110, Aug 2008.

[23] P. Bottorff and P. Saltsidis, "Scaling provider ethernet," *Communications Magazine, IEEE*, vol. 46, no. 9, pp. 104–109, September 2008.

[24] "IEEE Standard for Local and metropolitan area networks-Virtual Bridged Local Area Networks Amendment 10: Provider Backbone Bridge Traffic Engineering," *IEEE Std 802.1Qay-2009 (Amendment to IEEE Std 802.1Q-2005)*, pp. c1–131, Aug 2009.

[25] "IEEE Standard for Local and Metropolitan Area Networks - Virtual Bridged Local Area Networks Amendment 5: Connectivity Fault Management," *IEEE Std 802.1ag - 2007 (Amendment to IEEE Std 802.1Q - 2005 as amended by IEEE Std 802.1ad - 2005 and IEEE Std 802.1ak - 2007)*, pp. 1–260, 2007.

[26] M. Gupta and S. Singh, "Greening of the Internet," in *Proceedings of the 2003 Conference on Applications, Technologies, Architectures, and Protocols for Computer Communications*, ser. SIGCOMM '03. New York, NY, USA: ACM, 2003, pp. 19–26. [Online]. Available: http://doi.acm.org/10.1145/863955.863959

[27] K. J. Christensen, C. Gunaratne, B. Nordman, and A. George, "The next frontier for communications networks: power management," *Computer Communications*, vol. 27, no. 18, pp. 1758–1770, December 2004.

[28] A. Bianzino, C. Chaudet, D. Rossi, and J. Rougier, "A Survey of Green Networking Research," *Communications Surveys Tutorials, IEEE*, vol. 14, no. 1, pp. 3–20, First 2012.

[29] M. Gupta, S. Grover, and S. Singh, "A feasibility study for power management in LAN switches," in *Network Protocols, 2004. ICNP 2004. Proceedings of the 12th IEEE International Conference on*, Oct 2004, pp. 361–371.

[30] M. Gupta and S. Singh, "Using Low-Power Modes for Energy Conservation in Ethernet LANs," in *INFOCOM 2007. 26th IEEE International Conference on Computer Communications. IEEE*, May 2007, pp. 2451–2455.

[31] C. Gunaratne, K. Christensen, and B. Nordman, "Managing Energy Consumption Costs in Desktop PCs and LAN Switches with Proxying, Split TCP Connections, and Scaling of Link Speed," *Int. J. Netw. Manag.*, vol. 15, no. 5, pp. 297–310, Sept. 2005. [Online]. Available: http://dx.doi.org/10.1002/nem.565

[32] C. Gunaratne, K. Christensen, and S. Suen, "NGL02-2: Ethernet Adaptive Link Rate (ALR): Analysis of a Buffer Threshold Policy," in *Global Telecommunications Conference, 2006. GLOBECOM '06. IEEE*, Nov 2006, pp. 1–6.

[33] C. Gunaratne, K. Christensen, B. Nordman, and S. Suen, "Reducing the Energy Consumption of Ethernet with Adaptive Link Rate (ALR)," *Computers, IEEE Transactions on*, vol. 57, no. 4, pp. 448–461, April 2008.

[34] P. Purushothaman, M. Navada, R. Subramaniyan, C. Reardon, and A. George, "Power-Proxying on the NIC: A Case Study with the Gnutella File-Sharing Protocol," in *Local Computer Networks, Proceedings 2006 31st IEEE Conference on*, Nov 2006, pp. 519–520.

[35] K. Sabhanatarajan and A. Gordon-Ross, "A resource efficient content inspection system for next generation Smart NICs," in *Computer Design, 2008. ICCD 2008. IEEE International Conference on*, Oct 2008, pp. 156–163.

[36] S. Nedevschi, J. Chandrashekar, J. Liu, B. Nordman, S. Ratnasamy, and N. Taft, "Skilled in the Art of Being Idle: Reducing Energy Waste in Networked Systems," in *Proceedings of the 6th USENIX Symposium on Networked Systems Design and Implementation*, ser. NSDI'09. Berkeley, CA, USA: USENIX Association, 2009, pp. 381–394. [Online]. Available: http://dl.acm.org/citation.cfm?id=1558977.1559003

[37] J. Chabarek, J. Sommers, P. Barford, C. Estan, D. Tsiang, and S. Wright, "Power Awareness in Network Design and Routing," in *INFOCOM 2008. The 27th Conference on Computer Communications. IEEE*, April 2008, pp. –.

[38] B. Sanso and H. Mellah, "On reliability, performance and Internet power consumption," in *Design of Reliable Communication Networks, 2009. DRCN 2009. 7th International Workshop on*, Oct 2009, pp. 259–264.

[39] J. Blackburn and K. Christensen, "Green Telnet: Modifying a Client-Server Application to Save Eenrgy," *Dr. Dobb's Journal*, Oct 2008.

[40] J. Blackburn and K. Christensen, "A Simulation Study of a New Green BitTorrent," in *Communications Workshops, 2009. ICC Workshops 2009. IEEE International Conference on*, June 2009, pp. 1–6.

[41] L. Irish and K. Christensen, "A Green TCP/IP to reduce electricity consumed by computers," in *Southeastcon '98. Proceedings. IEEE*, Apr 1998, pp. 302–305.

[42] P. Reviriego, J. A. Hernández, D. Larrabeiti, and J. A. Maestro, "Burst transmission in Energy Efficient Ethernet," *IEEE Internet Computing*, vol. 14, no. 4, pp. 50–57, Jul/Aug. 2010.

[43] K. Christensen, P. Reviriego, B. Nordman, M. Benett, M. Mostowfi, and J. A. Maestro, "IEEE 802.3az: The road to Energy Efficient Ethernet," *IEEE Communications Magazine*, vol. 48, no. 11, pp. 50–56, 2010.

[44] M. Maier, *Optical Switching Networks*. New York: Cambridge University Press, 2008.

[45] R. Gaudino, "RINGO: demonstration of a WDM packet network architecture for metro applications," in *Transparent Optical Networks, 2002. Proceedings of the 2002 4th International Conference on*, vol. 1, 2002, pp. 77–80 vol.1.

[46] I. White, M. Rogge, K. Shrikhande, and L. G. Kazovsky, "A summary of the HORNET project: a next-generation metropolitan area network," *Selected Areas in Communications, IEEE Journal on*, vol. 21, no. 9, pp. 1478–1494, Nov 2003.

[47] N. Bouabdallah and H. Perros, "Cost-effective single-hub WDM ring networks: A proposal and analysis ," *Computer Networks*, vol. 51, no. 13, pp. 3878 – 3901, 2007. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1389128607001193

[48] H. Yin and D. Richardson, *Optical Code Division Multiple Access Communication Networks. Theory and Applications*. Beijing/New York: Springer, 2007.

[49] S. Maric, M. Hahm, and E. Titlebaum, "Construction and performance analysis of a new family of optical orthogonal codes for CDMA fiber-optic networks," *Communications, IEEE Transactions on*, vol. 43, no. 2/3/4, pp. 485–489, Feb 1995.

[50] G.-C. Yang and T. Fuja, "Optical orthogonal codes with unequal auto- and cross-correlation constraints," *Information Theory, IEEE Transactions on*, vol. 41, no. 1, pp. 96–106, Jan 1995.

[51] H. Chung and P. Kumar, "Optical orthogonal codes-new bounds and an optimal construction," *Information Theory, IEEE Transactions on*, vol. 36, no. 4, pp. 866–873, Jul 1990.

[52] Y. Chang, R. Fuji-Hara, and Y. Miao, "Combinatorial constructions of optimal optical orthogonal codes with weight 4," *Information Theory, IEEE Transactions on*, vol. 49, no. 5, pp. 1283–1292, May 2003.

[53] G.-C. Yang and W. C. Kwong, "Performance analysis of optical CDMA with prime codes," *Electronics Letters*, vol. 31, no. 7, pp. 569–570, Mar 1995.

[54] S. Maric, "New family of algebraically designed optical orthogonal codes for use in CDMA fibre-optic networks," *Electronics Letters*, vol. 29, no. 6, pp. 538–539, March 1993.

[55] S. Maric, Z. Kostic, and E. L. Titlebaum, "A new family of optical code sequences for use in spread-spectrum fiber-optic local area networks," *Communications, IEEE Transactions on*, vol. 41, no. 8, pp. 1217–1221, Aug 1993.

[56] S. Maric and E. L. Titlebaum, "A class of frequency hop codes with nearly ideal characteristics for use in multiple-access spread-spectrum communications and radar and sonar systems," *Communications, IEEE Transactions on*, vol. 40, no. 9, pp. 1442–1447, Sep 1992.

[57] M. Zheng and A. Albicki, "A modified hyperbolic congruential frequency hop codes for asynchronous event signaling," in *Computers and Communications, 1995., Conference Proceedings of the 1995 IEEE Fourteenth Annual International Phoenix Conference on*, Mar 1995, pp. 666–670.

[58] S. Maric and E. Titlebaum, "Frequency hop multiple access codes based upon the theory of cubic congruences," *Aerospace and Electronic Systems, IEEE Transactions on*, vol. 26, no. 6, pp. 1035–1039, Nov 1990.

[59] M. Kavehrad and D. Zaccarin, "Optical code-division-multiplexed systems based on spectral encoding of noncoherent sources," *Lightwave Technology, Journal of*, vol. 13, no. 3, pp. 534–545, Mar 1995.

[60] D. Zaccarin and M. Kavehrad, "An optical CDMA system based on spectral encoding of LED," *Photonics Technology Letters, IEEE*, vol. 5, no. 4, pp. 479–482, April 1993.

[61] Z. Wei, H. Shalaby, and H. Ghafouri-Shiraz, "Modified quadratic congruence codes for fiber Bragg-grating-based spectral-amplitude-coding optical CDMA systems," *Lightwave Technology, Journal of*, vol. 19, no. 9, pp. 1274–1281, Sep 2001.

[62] J. Salehi, A. Weiner, and J. Heritage, "Coherent ultrashort light pulse code-division multiple access communication systems," *Lightwave Technology, Journal of*, vol. 8, no. 3, pp. 478–491, Mar 1990.

[63] J. Cao, R. Broeke, C. Ji, Y. Du, N. Chubun, P. Bjeletich, T. Tekin, P. L. Stephan, F. Olsson, S. Lourdudoss, and S. Yoo, "Spectral encoding and decoding of monolithic InP OCDMA encoder," in *Optical Communication, 2005. ECOC 2005. 31st European Conference on*, vol. 3, Sept 2005, pp. 501–502 vol.3.

[64] C. Ji, R. Broeke, Y. Du, J. Cao, N. Chubun, P. Bjeletich, F. Olsson, S. Lourdudoss, R. Welty, C. Reinhardt, P. L. Stephan, and S. Yoo, "Monolithically integrated InP-based photonic chip development for O-CDMA systems," *Selected Topics in Quantum Electronics, IEEE Journal of*, vol. 11, no. 1, pp. 66–77, Jan 2005.

[65] V. Hernandez, W. Cong, J. Hu, C. Yang, N. Fontaine, R. Scott, Z. Ding, B. Kolner, J. Heritage, and S. Yoo, "A 320-Gb/s Capacity (32-User times; 10 Gb/s) SPECTS O-CDMA Network Testbed With Enhanced Spectral Efficiency Through Forward Error Correction," *Lightwave Technology, Journal of*, vol. 25, no. 1, pp. 79–86, Jan 2007.

[66] J. Heritage and A. Weiner, "Advances in Spectral Optical Code-Division Multiple-Access Communications," *Selected Topics in Quantum Electronics, IEEE Journal of*, vol. 13, no. 5, pp. 1351–1369, Sept 2007.

[67] N. Wada and K. Kitayama, "A 10 Gb/s optical code division multiplexing using 8-chip optical bipolar code and coherent detection," *Lightwave Technology, Journal of*, vol. 17, no. 10, pp. 1758–1765, Oct 1999.

[68] W. Huang, M. H. M. Nizam, I. Andonovic, and M. Tur, "Coherent optical CDMA (OCDMA) systems used for high-capacity optical fiber networks-system description, OTDMA comparison, and OCDMA/WDMA networking," *Lightwave Technology, Journal of*, vol. 18, no. 6, pp. 765–778, June 2000.

[69] K. Fouli and M. Maier, "OCDMA and Optical Coding: Principles, Applications, and Challenges [Topics in Optical Communications]," *Communications Magazine, IEEE*, vol. 45, no. 8, pp. 27–34, August 2007.

[70] S. Goldberg and P. Prucnal, "On the Teletraffic Capacity of Optical CDMA," *Communications, IEEE Transactions on*, vol. 55, no. 7, pp. 1334–1343, July 2007.

[71] J. Kani, K. Iwatsuki, and T. Imai, "Optical multiplexing technologies for access-area applications," *Selected Topics in Quantum Electronics, IEEE Journal of*, vol. 12, no. 4, pp. 661–668, July 2006.

[72] K. Kitayama, X. Wang, and N. Wada, "OCDMA over WDM PON-solution path to gigabit-symmetric FTTH," *Lightwave Technology, Journal of*, vol. 24, no. 4, pp. 1654–1662, April 2006.

[73] C.-S. Bres, I. Glesk, R. Runser, and P. Prucnal, "All optical OCDMA code drop unit for transparent ring networks," in *Lasers and Electro-Optics Society, 2004. LEOS 2004. The 17th Annual Meeting of the IEEE*, vol. 2, Nov 2004, pp. 501–502 Vol.2.

[74] C.-S. Bres, I. Glesk, R. Runser, and P. Prucnal, "All-optical OCDMA code-drop unit for transparent ring networks," *Photonics Technology Letters, IEEE*, vol. 17, no. 5, pp. 1088–1090, May 2005.

[75] Y. Deng, Z. Wang, K. Kravtsov, J. Chang, C. Hartzell, M. Fok, and P. Prucnal, "Demonstration and Analysis of Asynchronous and Survivable Optical CDMA Ring Networks," *Optical Communications and Networking, IEEE/OSA Journal of*, vol. 2, no. 4, pp. 159–165, April 2010.

[76] K. Kravtsov, Y. Deng, and P. Prucnal, "Self-Clocked All-Optical Add/Drop Multiplexer for Asynchronous CDMA Ring Networks," *Quantum Electronics, IEEE Journal of*, vol. 45, no. 4, pp. 396–401, April 2009.

[77] M. Gharaei, S. Cordette, I. Fsaifes, C. Lepers, and P. Gallion, "A novel ring architecture of multiple optical private networks over EPON using OCDMA technique," in *Transparent Optical Networks, 2009. ICTON '09. 11th International Conference on*, June 2009, pp. 1–4.

[78] B. Chen, C. Guo, J. Chen, L. Zhang, Q. Jiang, and S. He, "Add/drop multiplexing and TDM signal transmission in an optical CDMA ring network," *J. Opt. Netw.*, vol. 6, no. 8, pp. 969–974, Aug 2007. [Online]. Available: http://jon.osa.org/abstract.cfm?URI=jon-6-8-969

[79] T. Khattab and H. Alnuweiri, "Optical CDMA for All-Optical Sub-Wavelength Switching in Core GMPLS Networks," *Selected Areas in Communications, IEEE Journal on*, vol. 25, no. 5, pp. 905–921, June 2007.

[80] T. Khattab and H. Alnuweiri, "Optical GMPLS networks with code switch capable layer for sub-wavelength switching," in *Global Telecommunications Conference, 2004. GLOBECOM '04. IEEE*, vol. 3, Nov 2004, pp. 1786–1792 Vol.3.

[81] M. Meenakshi and I. Andonovic, "Code-based all optical routing using two-level coding," *Lightwave Technology, Journal of*, vol. 24, no. 4, pp. 1627–1637, April 2006.

[82] N. Calabretta, G. Contestabile, A. D'Errico, and E. Ciaramella, "All-optical label processing techniques for pure DPSK optical packets," *Selected Topics in Quantum Electronics, IEEE Journal of*, vol. 12, no. 4, pp. 686–696, July 2006.

[83] P. Polishuk, "Plastic optical fibers branch out," *IEEE Communications Magazine*, vol. 44, no. 9, pp. 140–148, Sept. 2006.

[84] Y. Koike, E. Nihei, N. Tanio, and Y. Ohtsuka, "Graded-index plastic optical fiber composed of methyl methacrylate and vinyl phenylacetate copolymers," *Appl. Opt.*, vol. 29, no. 18, pp. 2686–2691, Jun 1990. [Online]. Available: http://ao.osa.org/abstract.cfm?URI=ao-29-18-2686

[85] Y. Ohtsuka, E. Nihei, and Y. Koike, "Graded-index optical fibers of methyl methacrylate-vinyl benzoate copolymer with low loss and high bandwidth," *Applied Physics Letters*, vol. 57, no. 2, pp. 120–122, 1990. [Online]. Available: http://scitation.aip.org/content/aip/journal/apl/57/2/10.1063/1.103961

[86] I. Mollers, D. Jager, R. Gaudino, A. Nocivelli, H. Kragl, O. Ziemann, N. Weber, T. Koonen, C. Lezzi, B. A., and S. Randel, "Plastic optical fiber technology for reliable home networking: overview and results of the EU projet POF-ALL," *IEEE Communications Magazine*, vol. 47, no. 8, pp. 58–68, Aug. 2009.

[87] C. Okonkwo, E. Tangdiongga, H. Yang, D. Visani, S. Loquai, R. Kruglov, B. Charbonnier, M. Ouzzif, I. Greiss, O. Ziemann, R. Gaudino, and A. Koonen, "Recent Results From the EU POF-PLUS Project: Multi-Gigabit Transmission Over 1 mm Core Diameter Plastic Optical Fibers," *Lightwave Technology, Journal of*, vol. 29, no. 2, pp. 186–193, Jan 2011.

[88] *VDE 0885-763-1 standard, "Physical layer parameters and specification for high speed operation over Plastic Optical Fibres type HS-BASE-P"*, Oct. 2013.

[89] G. D. Forney, M. D. Trott, and S. Y. Chung, "Sphere-bound-achieving coset codes and multilevel coset codes," *IEEE Trans. Information Theory*, vol. 46, no. 3, pp. 820–850, May 2000.

[90] R. Sandberg, D. Goldberg, S. Kleiman, D. Walsh, and B. Lyon, "Design and Implementation of the Sun Network Filesystem," in *Proceedings of the USENIX 1985 Summer Conference*, June 1985, pp. 119–130.

[91] S. Shepler, M. Eisler, and D. Noveck, "Network File System (NFS) Version 4 Minor Version 1 Protocol." *IETF RFC 5661*, January 2010.

[92] M. Eisler, "RPCSEC GSS Version 2," *IETF RFC 5403*, February 2009.

[93] "Microsoft SMB Protocol and CIFS Protocol Overview (retreived May 2014)," *MSDN library, http://msdn.microsoft.com/en-us/library/windows/desktop/aa365233(v=vs.85).aspx*. [Online]. Available: http://msdn.microsoft.com/en-us/library/windows/desktop/aa365233(v=vs.85).aspx

[94] M. Chadalapaka, J. Satran, K. Meth, and D. Black, "Internet Small Computer System Interface (iSCSI) Protocol (Consolidated)," *IETF RFC 7143*, April 2014.

[95] J. Satran, K. Meth, C. Sapuntzakis, M. Chadalapaka, and E. Zeidner, "Internet Small Computer Systems Interface (iSCSI)," *IETF RFC 3720*, April 2004.

[96] "Open iSCSI project web page," *Open iSCSI, http://www.open-iscsi.org/*. [Online]. Available: http://www.open-iscsi.org/

[97] "Fibre Channel Framing and Signaling - 4 (FC-FS-4) Rev 0.40," *FC-FS-4 Standard, http://www.t11.org/ftp/t11/pub/fc/fs-4/14-018v0.pdf*, January 2014. [Online]. Available: http://www.t11.org/ftp/t11/pub/fc/fs-4/14-018v0.pdf

[98] "Fibre Channel Link Services (FC-LS-3) Rev 3.10," *FC-LS-3 Standard, http://www.t11.org/ftp/t11/pub/fc/ls-3/14-033v0.pdf*, February 2014. [Online]. Available: http://www.t11.org/ftp/t11/pub/fc/ls-3/14-033v0.pdf

[99] M. Rajagopal, E. Rodriguez, and R. Weber, "Fibre Channel Over TCP/IP (FCIP)," *IETF RFC 3821*, July 2004.

[100] C. Monia, R. Mullendore, F. Travostino, W. Jeong, and M. Edwards, "iFCP - A Protocol for Internet Fibre Channel Storage Networking," *IETF RFC 4172*, September 2005.

[101] "Fibre Channel - Fibre Channel Backbone - 5 (FC-BB-5)," *INCITS FC-BB-5*, 2009.

[102] C. S. white paper, "Fibre Channel over Ethernet Storage Networking Evolution," 2010, Cisco Systems White Paper.

[103] "IEEE Standard for Local and metropolitan area networks–Media Access Control (MAC) Bridges and Virtual Bridged Local Area Networks–Amendment 17: Priority-based Flow Control," *IEEE Std 802.1Qbb-2011*, pp. 1 –40, 30 2011.

[104] "IEEE Standard for Local and metropolitan area networks–Media Access Control (MAC) Bridges and Virtual Bridged Local Area Networks–Amendment 18: Enhanced Transmission Selection for Bandwidth Sharing Between Traffic Classes," *IEEE Std 802.1Qaz-2011*, pp. 1 –110, 30 2011.

[105] "IEEE Standard for Local and metropolitan area networks– Virtual Bridged Local Area Networks Amendment 13: Congestion Notification," *IEEE Std 802.1Qau-2010*, pp. c1 –119, 23 2010.

[106] S.-A. Reinemo, T. Skeie, and M. K. Wadekar, "Ethernet for high-performance data centers: On the new IEEE datacenter bridging standards," *IEEE Micro*, vol. 30, no. 4, pp. 42–51, 2010.

[107] C. S. white paper, "Priority Flow Control: Build reliable layer 2 infrastructure," 2009, Cisco Systems White Paper.

[108] M. Alizadeh, B. Atikoglu, A. Kabbani, A. Lakshmikantha, R. Pan, B. Prabhakar, and M. Seaman, "Data center transport mechanisms: Congestion control theory and IEEE standardization," in *Communication, Control, and Computing, 2008 46th Annual Allerton Conference on*, Sept 2008, pp. 1270–1277.

[109] B. Pranggono and J. Elmirghani, "A novel optical storage area network implemented in a metro WDM setting," in *High Performance Switching and Routing, 2005. HPSR. 2005 Workshop on*, May 2005, pp. 49–52.

[110] T. El-Gorashi, B. Pranggono, and J. Elmirghani, "Multi-Wavelength Metro WDM Sectioned Ring for SAN Extension under Hot Node Scenario and Variable Traffic Profiles," in *Transparent Optical Networks, 2006 International Conference on*, vol. 1, June 2006, pp. 139–146.

[111] T. El-Gorashi, B. Pranggono, and J. Elmirghani, "WDM Metropolitan Sectioned Ring for Storage Area Networks Extension with Symmetrical and Asymmetrical Traffic," in *Communications, 2006. ICC '06. IEEE International Conference on*, vol. 6, June 2006, pp. 2669–2674.

[112] B. Pranggono and J. M. Elmirghani, "Design and performance evaluation of a metro WDM storage area network with IP datagram support," *Optik - International Journal for Light and Electron Optics*, vol. 122, no. 18, pp. 1598 – 1602, 2011. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0030402610004845

[113] T. El-Gorashi, B. Pranggono, R. Mehmood, and J. Elmirghani, "A data Mirroring technique for SANs in a Metro WDM sectioned ring," in *Optical Network Design and Modeling, 2008. ONDM 2008. International Conference on*, March 2008, pp. 1–6.

[114] T. El-Gorashi and J. Elmirghani, "Distributed storage scenario in a wide area WDM mesh architecture under heterogeneous traffic," in *Optical Network Design and Modeling, 2009. ONDM 2009. International Conference on*, Feb 2009, pp. 1–6.

[115] S. Aiken, D. Grunwald, A. Pleszkun, and J. Willeke, "A performance analysis of the iSCSI protocol," in *Mass Storage Systems and Technologies, 2003. (MSST 2003). Proceedings. 20th IEEE/11th NASA Goddard Conference on*, April 2003, pp. 123–134.

[116] C. Gauger, M. Kohn, S. Gunreben, D. Sass, and S. Perez, "Modeling and performance evaluation of iSCSI storage area networks over TCP/IP-based MAN and WAN networks," in *Broadband Networks, 2005. BroadNets 2005. 2nd International Conference on*, Oct 2005, pp. 850–858 Vol. 2.

[117] Y. Zhang and M. MacGregor, "Tuning Open-iSCSI for Operation over WAN Links," in *Communication Networks and Services Research Conference (CNSR), 2011 Ninth Annual*, May 2011, pp. 85–92.

[118] G. Motwani and K. Gopinath, "Evaluation of advanced TCP stacks in the iSCSI environment using simulation model," in *Mass Storage Systems and Technologies, 2005. Proceedings. 22nd IEEE / 13th NASA Goddard Conference on*, April 2005, pp. 210–217.

[119] G. Motwani and K. Gopinath, "Evaluation of advanced TCP stacks in the iSCSI environment using simulation model," in *Mass Storage Systems and Technologies, 2005. Proceedings. 22nd IEEE / 13th NASA Goddard Conference on*, April 2005, pp. 210–217.

[120] M. Alizadeh, A. Greenberg, D. A. Maltz, J. Padhye, P. Patel, B. Prabhakar, S. Sengupta, and M. Sridharan, "Data Center TCP (DCTCP)," in *Proceedings of the ACM SIGCOMM 2010 Conference*, ser. SIGCOMM '10. New York, NY, USA: ACM, 2010, pp. 63–74. [Online]. Available: http://doi.acm.org/10.1145/1851182.1851192

[121] Q. Yang, "On Performance of Parallel iSCSI Protocol for Networked Storage Systems," in *Advanced Information Networking and Applications, 2006. AINA 2006. 20th International Conference on*, vol. 1, April 2006, pp. 629–636.

[122] B. Kancherla, G. Narayan, and K. Gopinath, "Performance Evaluation of Multiple TCP connections in iSCSI," in *Mass Storage Systems and Technologies, 2007. MSST 2007. 24th IEEE Conference on*, Sept 2007, pp. 239–244.

[123] F. Inoue, H. Ohsaki, Y. Nomoto, and M. Imase, "On Maximizing iSCSI Throughput Using Multiple Connections with Automatic Parallelism Tuning," in *Storage Network Architecture and Parallel I/Os, 2008. SNAPI '08. Fifth IEEE International Workshop on*, Sept 2008, pp. 11–16.

[124] H. Ohsaki, J. Kittananun, F. Inoue, Y. Nomoto, and M. Imase, "Performance evaluation of iSCSI-APT (iSCSI with Automatic Parallelism Tuning) on SINET3 with layer-1 bandwidth on demand service," in *Information and Telecommunication Technologies (APSITT), 2010 8th Asia-Pacific Symposium on*, June 2010, pp. 1–6.

[125] T. Nishijima, H. Ohsaki, Y. Nomoto, and M. Imase, "Performance Evaluation of Block Device Layer with Automatic Parallelism Tuning Using Heterogeneous IP-SAN Protocols," in *Applications and the Internet (SAINT), 2010 10th IEEE/IPSJ International Symposium on*, July 2010, pp. 343–346.

[126] A. Reid, P. Willis, I. Hawkins, and C. Bilton, "Carrier ethernet," *Communications Magazine, IEEE*, vol. 46, no. 9, pp. 96–103, September 2008.

[127] M. Ali, G. Chiruvolu, and A. Ge, "Traffic engineering in Metro Ethernet," *Network, IEEE*, vol. 19, no. 2, pp. 10–17, March 2005.

[128] Y. Gao, Y. Wang, W. Ma, J. Wu, and J. Lin, "Code-division multiple-access in an optical fiber LAN with ring," in *Communication Technology Proceedings, 2003. ICCT 2003. International Conference on*, vol. 1, April 2003, pp. 743–747 vol.1.

[129] J. Wu and C.-L. Lin, "Fiber-optic code division add-drop multiplexers," *Lightwave Technology, Journal of*, vol. 18, no. 6, pp. 819–824, June 2000.

[130] R. Gordon and L. Chen, "Demonstration of all-photonic code conversion in a semiconductor fiber ring laser for OCDMA networks," in *Information Photonics, 2005. IP 2005. OSA Topical Meeting on*, June 2005, pp. 1–3.

[131] A. Tychopoulos, O. Koufopavlou, and I. Tomkos, "FEC in optical communications - A tutorial overview on the evolution of architectures and the future prospects of out-band and inband FEC for optical communications," *Circuits and Devices Magazine, IEEE*, vol. 22, no. 6, pp. 79–86, Nov 2006.

[132] T. El-Gorashi, A. Mujtaba, W. Adlan, and J. Elmirghani, "Storage area networks extension scenarios in a wide area WDM mesh architecture under heterogeneous traffic," in *ICTON '09. 11th Int. Conf. Transparent Optical Networks (ICTON)*, July 2009.

[133] T. El-Gorashi and J. Elmirghani, "Data replication schemes for a distributed storage scenario," in *ICTON '10. 12th Int. Conf. on Transparent Optical Networks (ICTON)*, July 2010.

[134] J. Cordeiro and J. Kharoufeh, "Batch Markovian Arrival Processes (BMAP)," 2011, (Technical report). [Online]. Available: http://www.pitt.edu/~jkharouf/Papers/Cordeiro_Khar_BMAP_Final.pdf

[135] A. Grzemba, *MOST - The automitve multimedia network from MOST25 to MOST150*, 2008.

[136] *IEEE Std 802.3, Gigabit Ethernet over Plastic Optical Fiber Study Group*, 2014.

[137] M. Jinly, R. Fengyuan, J. Wanchun, and L. Chuang, "Modeling and understanding burst transmission algorithms for Energy Efficient Ethernet," in *Proc. of the IEEE/ACM 21st Int. Symp. on Quality of Service (IWQoS)*, June 2013.

[138] S. Herrería-Alonso, M. Rodríguez-Pérez, M. Fernández-Veiga, and C. López-García, "Bounded energy consumption with dynamic packet coalescing," in *Proc. of IEEE NOC*, June 2012.

[139] A. Chatzipapas and V. Mancuso, "Modelling and real-trace-based evaluation of static and dynamic coalescing for Energy Efficient Ethernet," in *Proc. of Int. Conf. on Future Energy Systems (e-Energy)*, 2013, pp. 161–172.

[140] D. Fedyk and D. Allan, "Ethernet data plane evolution for provider networks [next-generation carrier ethernet transport technologies]," *Communications Magazine, IEEE*, vol. 46, no. 3, pp. 84–89, March 2008.

[141] F. Chung, J. Salehi, and V. Wei, "Optical orthogonal codes: design, analysis and applications," *Information Theory, IEEE Transactions on*, vol. 35, no. 3, pp. 595–604, May 1989.

[142] J. Baliga, R. Ayre, K. Hinton, W. Sorin, and R. Tucker, "Energy Consumption in Optical IP Networks," *Lightwave Technology, Journal of*, vol. 27, no. 13, pp. 2391–2403, July 2009.

[143] *CAIDA Anonymized Internet Traces 2009 Dataset*.

Acronyms

AIS *Alarm Indication Signal*

API *Application Programming Interface*

ATM *Asynchronous Transfer Mode*

B-DA *Backbone Destination Address*

B-SA *Backbone Source Address*

B-Tag *Backbone VLAN Tag (B-VID)*

B-VID *Backbone VLAN ID*

B-VLAN *Backbone VLAN*

BCB *Backbone Core Bridge*

BCH *Bose-Chaudhuri-Hocquenghem*

BEB *Backbone Edge Bridge*

BER *Bit Error Rate*

BPDU *Bridge Protocol Data Unit*

BPP *Burst Poisson Process*

BSI *Backbone Service Instance*

CBS *Commited Burst Size*

CCC *Cubic Congruence Code*

CCM *Continuity Check Message*

CDF *Cumulative Distribution Function*

CDMA *Code Division Multiple Access*

CEE *Convergence Enhanced Ethernet*

CFM *Connectivity Fault Management*

CIR *Commited Information Rate*

CN *Congestion Notification*

CNM *Congestion Notification Message*

CSMA/CA *Carrier Sense Multiple Access/Collision Avoidance*

CSMA/CD *Carrier Sense Multiple Access/Collision Detection*

CST *Common Spanning Tree*

CTMC *Continuous Time Markov Chain*

DBORN *Dual Bus Optical Ring Network*

DCB *Data Centre Bridging*

DFE *Decision Feedback Equalisation*

EBS *Excess Burst Size*

EDFA *Erbium Doped Fibre Amplifier*

EEE *Energy Efficient Ethernet*

EIR *Excess Information Rate*

EO *Electrical to Optical*

EPON *Ethernet PON*

ESP *Ethernet Switched Path*

ESP-MAC DA *ESP-MAC Destination Address*

ESP-MAC SA *ESP-MAC Destination Address*

ESP-VID *ESP VLAN ID*

ETS *Enhanced Transmission Selection*

EVC *Ethernet Virtual Connection*

FBG *Fibre Bragg Grating*

FC *Fibre Channel*

FC-BB-5 *Fibre Channel Backbone 5*

FCIP *Fibre Channel over IP*

FCoE *Fibre Channel over Ethernet*

FCP *Fibre Channel Protocol*

FDMA *Frequency Division Multiple Access*

FEC *Forward Error Correction*

GMII *Gigabit MII*

GMPLS *Generalised MPLS*

HBA *Host Bus Adapter*

HCC *Hyperbolic Congruence Code*

HORNET *Hybrid Optoelectronic Ring Network*

I-SID *PBB Backbone Service Instance Identifier*

iFCP *Internet Fibre Channel Protocol*

IPTV *IP television*

IS-IS *Intermediate System to Intermediate System routing protocol*

iSCSI *Internet SCSI*

LAN *Local Area Network*

LBM *Loopback Message*

LBR *Loopback Reply*

LPI *Low Power Idle*

LTR *Linktrace Message*

LTR *Linktrace Reply*

LUN *Logical Unit Number*

MAC *Medium Access Control*

MAI *Multiple Access Interference*

MAN *Metropolitan Area Network*

MAWSON *Metropolitan Area Wavelength Switched Optical Network*

ME *Metro Ethernet*

MII *Media Independent Interface*

MLCC *Multilevel Coset Coding*

MPLS *Multi Protocol Label Switching*

MSTI *Multiple Spanning Tree Instance*

MSTP *Multiple Spanning Tree Protocol*

NAS *Network Attached Storage*

NFS *Network File System*

NIC *Network Interface Card*

OAM *Operation And Management*

OBS *Optical Burst Switching*

OCDMA *Optical Code Division Multiple Access*

OCS *Optical Circuit Switching*

OE *Optical to Electrical*

OLT *Optical Line Terminal*

ONU *Optical Network Unit*

OOC *Optical Orthogonal Code*

PB *Provider Bridges*

PBB *Provider Backbone Bridges*

PBB-TE *Provider Backbone Bridges Traffic Engineering*

PDU *Protocol Data Unit*

PFC *Priority Flow Control*

PHS *Physical Layer Header*

PMMA POF *Poly Methyl Methacrylate POF*

PMMA-GI POF *PMMA Graded Index POF*

POE *Power Over Ethernet*

POF *Plastic Optical Fibre*

PON *Passive Optical Network*

QCC *Quadratic Congruence Code*

RAID *Redudant Array of Independent Disks*

RINGO *Ring Optical Network*

RPS *Rapid PHY Selection*

RSTP *Rapid Spanning Tree Protocol*

RTSP *Rapid Spanning Tree Protocol*

RTT *Round Trip Time*

SAE *Spectral Amplitude Encoding*

SAN *Storage Area Network*

SCSI *Small Computer System Interface*

SI-PMMA-POF *Step Index PMMA POF*

SLA *Service Level Agreement*

SMB/CIFS *Server Message Block/Common Internet File System*

SONET/SDH *Synchronous Optical Networking/Synchronous Digital Hierarchy*

SPB *Shortest Path Bridging*

SPE *Spectral Phase Encoding*

SSMF *Standard Single Mode Fibre*

STP *Spanning Tree Protocol*

TDM *Time Division Multiplex*

TDMA *Time Division Multiple Access*

TE-SID *Traffic Engineering Service Instance Identifier*

TESI *Traffic Engineering Service Instance*

TPE *Temporal Phase Encoding*

TRILL *Transparent Interconnection of Lots of Links*

TT-FR *Tunable Transmitter Fixed Receiver*

UNI *User Network Interface*

VCSEL *Vertical-Cavity Surface Emitting Laser*

VLAN *Virtual Local Area Network*

VOD *Video On Demand*

VPLS *Virtual Private Line Services*

WAN *Wide Area Network*

WDM *Wavelength Division Multiplex*

WDMA *Wavelength Division Multiple Access*

WRN *Wavelength-Routed Network*

XGMII *Ten Gigabit MII*